scottish institute for research in economics

SIRE DISCUSSION PAPER

SIRE-DP-2013-75

Intergenerational Mobility and the Informative Content of Surnames

Maia Güell
The University of Edinburgh, CEP (LSE), CEPR & IZA

José V. Rodríguez Mora
The University of Edinburgh and CEPR

Christopher I. Telmer
Carnegie Mellon University

www.sire.ac.uk

Intergenerational Mobility and the Informative Content of Surnames*

Maia Güell
University of Edinburgh,
CEP (LSE), CEPR & IZA

José V. Rodríguez Mora†
University of Edinburgh and CEPR

Christopher I. Telmer
Tepper School of Business
Carnegie Mellon University

First version: May 2007
This version: October 2013

Abstract

We propose a new methodology for measuring intergenerational mobility in economic well-being. Our method is based on the joint distribution of *surnames* and economic outcomes. It circumvents the need for intergenerational panel data, a long-standing stumbling block for understanding mobility. A single cross-sectional dataset is sufficient. Our main idea is simple. If 'inheritance' is important for economic outcomes, then *rare* surnames should predict economic outcomes in the cross-section. This is because rare surnames are indicative of familial linkages. Of course, if the number of rare surnames is small, this won't work. But rare surnames are abundant in the highly-skewed nature of surname distributions from most Western societies. We develop a model that articulates this idea and shows that the more important is inheritance, the more informative will be surnames. This result is robust to a variety of different assumptions about fertility and mating. We apply our method using the 2001 census from Catalonia, a large region of Spain. We use educational attainment as a proxy for overall economic well-being. Our main finding is that mobility has *decreased* among the different generations of the 20th century. A complementary analysis based on sibling correlations confirms our results and provides a robustness check on our method. Our model and our data allow us to examine one possible explanation for the observed decrease in mobility. We find that the degree of *assortative mating* has increased over time. Overall, we argue that our method has promise because it can tap the vast mines of census data that are available in a heretofore unexploited manner.

Key words: Surnames, intergenerational mobility, cross-sectional data analysis, population genetics, assortative mating, siblings.

JEL codes: C31, E24, J1

*We thank Laurence Ales, Namkee Ann, Manuel F. Bagüés, Melvin Coles, Vicente Cuñat, John Hassler, John Knowles, Ramon Marimon, Laura Mayoral, John Moore, Diego Puga, Gary Solon, Murat Tasci and numerous seminar participants for very useful suggestions, and Anisha Gosh, Rasa Karapanza, Ana Mosterín and Robert Zymek for superb assistance. Financial support from the Spanish Ministry of Education and Science under grants SEJ2006-09993 (MG) and SEJ2007-64340 (JVRM) is gratefully acknowledged. Our online appendix is available at http://www.sevirodriguezmora.com/grt/GRT_online_appendix.pdf

†Corresponding author: School of Economics, University of Edinburgh, 31 Buccleuch Place, Edinburgh EH8 9JT, United Kingdom. Email: sevimora@gmail.com

1 Introduction

The empirical challenges to understanding intergenerational mobility are large. The main reasons are data limitations. To address the issue directly, panel data is needed that links the economic status of adults to that of their parents for multiple generations. The existence of such data is rare. That which does exist (i) has been strongly criticized in terms of various biases, and (ii) cannot address the question of how mobility has *changed* over time.[1] This last point deserves emphasis. The complexity of intergenerational mobility makes its measurement at a point in time very difficult to do and interpret. Comparing measures of mobility across time would be far less problematic ... if we had the data.

This paper attempts to make headway by introducing a new source of data: *surnames* and how they vary with measures of economic well-being. We show that the implicit intergenerational links that are inherent in surnames provide a useful stand-in for the explicit intergenerational links that would exist in multi-generational panel data. We do so using both theory and data. Our theory shows how surname data can allow one to estimate both the level and the change in intergenerational mobility *even in the absence of explicit links between children and their parents*. Our empirical work implements this idea. We use novel, census-based data from Catalonia — a large region of Spain — to obtain measures of mobility both at one point in time and across generations. The former concord well with previous studies. The latter, more strikingly, provide an estimate of how mobility has *changed* over time. We find that, among the different generations of the 20th century, it has *decreased*. That is, the economic status of parents seems to have become *more* closely related to that of their children. Why? Our methodology offers one potential explanation. Assortative mating — the tendency for people with similar economic status to mate with one another — has increased over time. We use our surname data to establish this and our model to demonstrate how it is one potential source of a decrease in intergenerational mobility.

Our main idea is straightforward. Intergenerational mobility is all about how children inherit aspects of economic well-being from their parents. In data, we can observe *outcomes* on economic well-being, but, unless the data contain explicit knowledge of who is the parent of who, we are unable to determine the degree to which the economic well-being has been inherited. However, suppose that the data also allow us to observe surnames. Surnames are almost always inherited

[1] See Solon (1992), Haider and Solon (2006), Hertz (2007) and Lee and Solon (2009). Extensive literature surveys are available in Solon (1999) and Black and Devereux (2011).

from one's father. Thus, they can serve as *markers*. They are intrinsically irrelevant for the determination of economic well-being, but they get passed from one generation to the next, alongside other characteristics that *do* matter. The more important are these characteristics in determining outcomes, the more inheritable are the outcomes, and, therefore, the more information the surnames will contain on the values of outcomes. In this way, surnames can be used to measure the importance of inheritance and thus identify the degree of mobility. The following example articulates this mechanism in more detail.

Consider a society comprised of two distinct groups of people: rich and poor. In each group there are males and females, but, because surnames are (typically) passed along the male lineage, we will ignore the females for now. Suppose that the males within each group all share the same surname: Richmanson for the rich and Poormanson for the poor. This means that if we partition society either by surname or by economic status, we get the same thing. We would say that, among this initial generation, surnames are 'perfectly informative.' If you know a man's surname, you know his economic status. Now, how informative will surnames be among the subsequent generation? The answer depends on the degree of inheritance. Consider two extremes. First, if inheritance is 'perfect' — meaning that there is no mobility whatsoever — then the economic status of all sons would be identical to that of their fathers. Surnames, being passed from father to son in exactly the same manner as economic status, would remain perfectly informative. Second, if inheritance is irrelevant, so that the sons of both the Richmansons and Poormansons are equally likely to be rich or poor, then surnames would become perfectly *un*informative among the next generation. Thus, the informativeness of surnames depends on the degree of intergenerational mobility. This is the essence of the mechanism with which we are able to infer the degree of mobility.

Reality, of course, lies between these two extremes, and surnames carry some information. But, there is a tendency for surnames to become non-informative as time goes on. To understand this, suppose that economic status follows a stationary process with some degree of persistence, and that this process is the same for both the Richmansons and the Poormansons. In this case, surnames will remain informative among the sons of the initial generation, but the informativeness will not be perfect. It will become less perfect with each subsequent generation. After enough generations the cross-sectional distribution of status among both elements of the surname partition will be the same and the initial informativeness will have vanished. This is what makes our methodology somewhat

less than obvious. For surnames to contain information about intergenerational mobility there must be something else going on that inhibits this convergence to a common stationary distribution.

The additional ingredient in our study is a *birth-death process* for surnames. Surnames die when the last male holder of a particular name bears no male children. They are born when someone mutates their name, or when a new name enters the population via immigration. This generates a *skewed* surname distribution, with a small number of names each being held by a large number of people and a large number of names each being held by a small number of people. The surname distribution in most Western societies takes exactly this form.

Herein lies the key to our method: skewness in the surname distribution. We can't learn anything from the name Smith. The cross-sectional distribution among the Smiths is similar to that among society as a whole. We can, however, learn something from the multitude of *rare* surnames. This is because they form a partition of the population that is correlated with familial linkages. Suppose, for example, that a rich person chooses a new, unique surname and that this person's male descendants maintain this name. Then, if mobility is low, this surname will be informative for some number of subsequent generations. It will be shared by a bunch of rich people. If mobility is high it will not be as informative. The same logic holds for a poor person who changes their name, or for an immigrant with a distinct name. Over time these people's surnames will either die or will multiply. If they multiply, then, as discussed above, they are likely to become less informative. But, overall, a stationary birth/death process will generate a stationary *commonality-ordered* surname frequency distribution. There will always be some rare names. These are the names from which we can extract information on intergenerational mobility.

An important issue for us is ethnicity. Early-generation immigrants, for example, tend to have distinctive surnames and relatively low economic status. Surnames, therefore, contain information on both familial linkages *and* ethnicity. Ethnicity, therefore, is likely to make surnames economically informative, in-and-of-itself. One might even think that the *only* reason that surnames are informative is because they reveal ethnicity. We show that this is not the case. Our data permit us to control for ethnicity and we find that, once we do, surnames remain informative. We provide supplemental evidence indicating that this is because our methodology is able to reveal familial linkages when applied to ethnically-homogeneous populations. Our method seems to be measuring what our theory indicates that it should be measuring. In Section 5 we relate these findings to

existing papers on intergenerational mobility, a literature that typically *amalgamates* the inheritance of ethnicity with the inheritance of unobservable economic traits. Our approach allows us to separately identify these things. We report parameter estimates that both control for ethnicity for methodological reasons and, in order to draw links to the existing literature, estimates that do not.

Our method is not without its limitations, but they are counterbalanced by some important strengths. One is that it is quite data-friendly. We are able to measure mobility from a single, cross-sectional census. We do not require any explicit links between parents and children. As a result, the censuses that are periodically compiled by many governments contain most of the information that we require. Moreover, a great deal of confidentiality and anonymity can be maintained while still allowing access to the necessary information. Surnames can be encoded without negating what we do.

Our method can be applied to any country that follows the Western naming convention. Spain, the source of our data, uses a variant of the Western convention that is *identical* to its Anglo-Saxon counterpart, but with the additional ingredient of the maternal surname, which survives for one generation. We show that knowledge of the maternal name can be exploited in three ways: (i) to control for ethnicity, (ii) to determine the degree of assortative mating among the parents of an individual, and (iii) to partition our data into sets of siblings. The latter affords us a powerful robustness check on our methodology.

Existing literature that is closely related to our study is as follows. First, a number of papers have studied the distribution of *first names* to understand phenomenon ranging from racial discrimination and economic status (e.g., Bertrand and Mullainathan (2004), Fryer and Levitt (2004), Levitt and Dubner (2005)) to mobility (Olivetti and Paserman (2011)). *Endogeneity* plays a key role here. Parents *choose* first names and these choices can be related to parental characteristics. Our study, in sharp contrast, makes use of the fact that surnames are surely much more *exogenous* in nature, playing the role of innocuous markers. Second, papers that have used *last names* to study economic phenomenon include Angelucci, De Giorgi, Rangel, and Rasul (2010), Bagüés (2005), Collado, Ortuño-Ortín, and Romeo (2012a), Collado, Ortuño-Ortín, and Romeo (2012b), Long and Ferrie (2011). These works use surnames as family links explicitly and intensively (i.e., to determine familial links in a small sample) while we use them implicitly and extensively for the

whole population. Third, a large literature on mobility has devised many creative ways to overcome limited panel data availability. Examples include Aaronson and Mazumder (2008), Dahan and Gaviria (2001), Duncan, Featherman, and Duncan (1972), Levine and Mazumder (2007), Page and Solon (2003), and Solon, Corcoran, Roger, and Deborah (1991). Again, most of these papers rely on explicit familial linkages, the lack of which is a distinguishing feature of our approach. Finally, most closely related to us is Clark (2013), who also uses rare surnames to study intergenerational mobility. What distinguishes our work is that (i) it predates his, being first published as 2007 CEPR Discussion Paper #6316, (ii) we use the complete census data which is absent from selection issues, and (iii) we use an explicit model to map characteristics of the surname distribution into the intergenerational correlations that are typical in the literature. To our knowledge, our paper is the first to use surnames to study intergenerational mobility in this manner.

The remainder of our paper is organized as follows. In Section 2 we define an empirical measure of the extent to which surnames are informative for the economic status of their owners. We label this measure the Informational Content of Surnames (ICS). Section 3 develops and analyzes a model of the joint distribution of surnames and economic outcomes. This model maps the ICS into an intergenerational correlation that is standard in the literature. Section 4 describes the census data that we use to implement our model. Section 5 makes precise how we control for ethnicity and how the parameters that we try to estimate relate to those that typify the literature. Section 6 reports results on mobility in Catalonia in the year 2001, and Section 7 uses these results to calibrate our model. Section 8 reports results on how mobility has changed over time. Section 9 affirms the robustness of our methodology by constructing and analyzing an alternative dataset based on siblings. Section 10 presents additional theory and evidence suggesting that changes in assortative mating may have played an important role. Section 11 concludes.

2 The Informational Content of Surnames

The population consists of N individuals. Each individual is associated with one surname, s, which is an element of the finite set of all possible surnames, Ω. A *census* is list with one entry per individual in the population. The i^{th} entry records individual i's surname, a measure of their economic well-being, e_{is}, and a vector of additional characteristics, X_{is}, such as age, gender, ethnicity, place of birth, etc. We model economic inheritance as being described by

$$e_{is} = \gamma' X_{is} + y_{is} \tag{1}$$

$$y_{is} = \rho\, y_{ip} + \varepsilon_{is} \tag{2}$$

where γ is a vector of parameters, ε_{is} is an *iid* shock with variance V_ε and y_{is} is a set of unobservable traits that are passed from parents, who have traits y_{ip}, to their children. The parameter $0 \leq \rho < 1$ measures the importance of economic inheritance.

We define the *informational content of surnames* (ICS) as the difference in the R^2 between two regressions. The first, with R^2 denoted R^2_L, estimates equation (1) for the *average* individual with surname s

$$e_{is} = \gamma' X_{is} + b' D + \text{residual} \ , \tag{3}$$

where D is an S-vector of surname-dummy variables with $D_s = 1$ if individual i has surname s and $D_s = 0$ otherwise. Our methodology is based on the idea that as surname s becomes more infrequent it becomes more likely that this average gets taken across individuals with familial linkages, thereby providing information about economic inheritance.

The second regression mixes up the surnames so that they cannot be informative. It is

$$e_{is} = \gamma' X_{is} + b' F + \text{residual} \ , \tag{4}$$

where F is an S-vector of 'fake' dummy variables that randomly assign surnames to individuals in a manner that maintains the marginal distribution of surnames. The R^2 from this regression is denoted R^2_F. The ICS is defined as

$$\text{ICS} \equiv R^2_L - R^2_F \ . \tag{5}$$

The ICS is a moment of the joint distribution of surnames and economic well-being that measures the *incremental* informational content of surnames. In our model it will turn out to be monotonically increasing in the economic inheritance parameter, ρ.

The basic idea behind the ICS measure is this. Surnames define a *partition* of the population. If the surname partition is informative about familial linkages — if some individuals with the same surname come from the same family — then it can be used to measure the importance of economic inheritance. The fake-surname partition *is constructed* to have zero information about familial linkages. By comparing the relative informativeness of the two partitions, therefore, we measure

the extent to which surnames contain incremental information.

The ICS measure has a number of important advantages. Suppose, for example, that every individual had a unique surname. Then, by definition, surnames would contain zero information. In this case $R_L^2 = 1$ but ICS=0 indicating, as it should, that surnames contain no information. Moreover, the number of surnames in a typical census is large and grows with the population size so that the D matrix from Equation (3) has many columns. The ICS measure insures that any information that exists as a consequence of the large partition of the population used (many dummies) is not attributed to surnames.

In Section 6.1 we estimate the ICS using Spanish census data. Before doing so, we develop a model that allows us to map ICS units into 'inheritance units:' an intergenerational correlation coefficient.

3 Model

In Western societies the intergenerational transmission of surnames occurs primarily between fathers and sons. We therefore begin by ignoring females. At date $t-1$ the population consists of N_{t-1} males. Each individual reproduces with probability q. Conditional on reproducing, an individual gives birth to m sons. Generations do not overlap. Fathers die after reproducing (or failing to reproduce). The expected growth rate of the population is $mq - 1$, which we assume to be zero.

Each of the N_{t-1} individuals is associated with one surname from the fixed, discrete set Ω. The typical element of Ω, denoted $s \in \Omega$, can take on one of S different values (so that $S \equiv \#\Omega$). The date $t-1$ marginal distribution of surnames is $F_{t-1} : \Omega \to [0,1]$. The number of *active* surnames at date $t-1$, denoted $S_{t-1} < S$, is therefore equal to the number of strictly positive values of $F_{t-1}(s)$. If each individual has a unique surname then $S_{t-1} = N_{t-1}$. Otherwise $S_{t-1} < N_{t-1}$. In reality — and in our model — N_{t-1} is far greater than S_{t-1}. The initial distribution is denoted F_0.

The surname distribution evolves from F_{t-1} to F_t according to a birth/death process. The death of a surname occurs if all fathers possessing that name bear zero offspring.[2] Birth occurs via *mutation*: a son acquiring a different (typically new) surname than his father. Mutations are a necessary ingredient of our methodology. As we show in Section 3.1, without them the

[2]The death of surname s can also occur if all sons of all fathers with surname s mutate their names. Quantitatively, however, the likelihood of this happening is dwarfed by the likelihood that these fathers simply give birth to zero sons.

surname distribution would neither be informative nor would it resemble the highly skewed nature of observed surname distributions. Note that there is a certain irony here. Mutations seemingly frustrate the surname researcher: they 'destroy' intergenerational linkages. Yet without them the surname researcher would eventually be out of business.

Formally, mutation occurs as follows. At date t each existing name, $s : F_{t-1}(s) > 0$, will have vanished, so that $F_t(s) = 0$, if all fathers possessing that name at $(t-1)$ bear zero offspring. This occurs with probability $(1-q)^{N_s}$ where $N_s = F_{t-1}(s)N_{t-1}$, the number of fathers with name s. A surviving surname matches that of the father with probability $(1-\mu)$ and *mutates* with probability μ. A mutated surname is simply a new name, $s \in \Omega$, chosen randomly.

Economic well-being is passed from fathers to sons according to equations (1)–(2) with $\gamma = 0$. That is, we ignore cross-sectional variation in the X_{is} directions so that individuals differ only in terms of surname and economic inheritance, y_{is}. Economic inheritance and well-being are therefore the same thing, $e_{is} = y_{is}$, which we refer to as *income* for simplicity. Rewriting equation (2), we have that the income of individual i with surname s at date t is determined by

$$y_{ist} \quad = \quad \rho\, y_{ip,t-1} + \varepsilon_{ist} \quad , \tag{6}$$

where $y_{ip,t-1}$ is individual i's father's income, one generation removed. By definition the surname associated with $y_{ip,t-1}$ is the same as that of y_{ist} (unless mutation occurs). Note that the mean of y_{ist} is zero, meaning that we are dealing with demeaned data relative to what is implicit in X_{is} from equation (1).

Siblings are individuals with y_{ist} and y_{jst} such that their surname s and parent p are the same (again, ignoring mutation). Identical surnames can also be associated with cousins, second cousins and so on. On the other hand, identical surnames can arise purely by chance, in the absence of any familial linkages. The smaller is $F_{t-1}(s)$, the less likely this is (*e.g.*, if $N_s = 1$ it is impossible).

The sense in which ρ relates to families and inheritance is manifest in the the distinction between the conditional and the unconditional variance of y_{ist}. For example, the cross-sectional variance between siblings is equal to the conditional variance, V_ε. For cousins it is $V_\varepsilon(1+\rho^2)$. For the entire population it coincides with the unconditional variance, $V_\varepsilon/(1-\rho^2)$. A larger inheritance parameter, ρ, therefore implies lower cross-sectional variance between family members *relative to* overall cross-sectional variance. The larger is ρ the larger will be the tendency for a surname to link two people with similar incomes, *relative to* two people randomly chosen from the population

with, typically, different surnames.

3.1 Analysis

We now discuss the key features of our model. Some features can be characterized analytically while for others we must rely on simulation-based evidence. For the simulations we use the following baseline parameter values. First, we abstract from growth so that the expected population growth rate, $mq - 1$ is zero. To achieve this, we set the reproduction probability to $q = 1/2$ and the number of offspring to $m = 2$.

Second, we choose the conditional variance, $V_\varepsilon = 1$, and the mutation rate, $\mu = 0.02$. Third, the initial number of individuals, N_0, is set to 1 million and the initial surname and income distributions are uniform. Finally, we vary the inheritance parameter, ρ, from 0.05 to 0.95. Whenever appropriate we examine the sensitivity of our results to departures from these baseline values.

The most important feature of our model is that skewness in the surname distribution gives rise to the informational content of surnames, and that this informational content is increasing in the inheritance parameter, ρ. We demonstrate this in the following sequence of properties.

Property 1 : Random walk behavior

Suppose that there is no surname mutation, $\mu = 0$, and the expected growth rate of the population is zero, $mq = 1$. Then the number of individuals with surname $s \in \Omega$ follows a driftless random walk with an absorbing barrier at zero.

The proof is in Appendix A. It is a simple consequence of the fact that the number of individuals with a given surname is a binomial random variable. Its importance is that it tells us that mutation is *necessary* for surnames to be informative. This is because a driftless random walk will, given enough time, visit all parts of its sample space. Eventually, therefore, all but one (non-informative) surname will disappear prior to the disappearance of the population (which, with $mq = 1$, must also eventually happen).

Next we demonstrate the way in which mutation generates skewness and thus informativeness. To do so we work with the ordered frequency distribution, denoted $G_t : [1, 2, \ldots, S] \to [0, 1]$. This distribution simply provides the relative frequency of the most common surname, the second most common surname, and so on. The long-run distribution associated with $G_t(k)$ is denoted $G(k)$, for $k = 1, 2, \ldots, S$.[3]

[3]Formally, order the elements of Ω (arbitrarily) so that we can write $\Omega = \{s_1, s_2, \ldots, s_S\}$. Define the ranking

Property 2 : Skewed surname distribution

Given zero expected population growth, $mq = 1$, and a mutation rate, $0 < \mu < 1$, then for any initial distribution, $F_0(s)$ (and the associated $G_0(k)$), there exists a $k > 0$ such that, for all $t > k$, the distributions $G_t(k)$ display three key properties: (i) they are highly skewed, (ii) the number of individuals per surname is a constant, (iii) the Gini coefficient is a constant.

Figure 1 plots time-series, $t = 1$ to $t = 400$, of the number of individuals per surname and the Gini coefficient of the distributions $G_t(k)$. In each graph there are four time series, each one corresponding to a different initial condition for the number of surnames (described in the caption). These moments of the distribution have clearly converged, thus validating Property 2. Since the Pareto distribution is completely characterized by these two moments, it seems likely that $G_t(k)$, also plotted in the figure for $t > k$ (along with the associated Lorenz curve), is a Pareto distribution. For our purposes, however, the exact form of $G_t(k)$ is not critical. What is critical is the behavior of the ICS, discussed below.

The skewness in Figure 1 is what drives our methodology. To understand why, consider first the names that occur with a high frequency. Since the income process in equation (6) is stationary, the cross-sectional distribution among these names is very similar to that of the overall population. These names therefore cannot be informative. In contrast, consider the very infrequent names. Many of them derive from recent mutations. They are newly created names, or the names of sons of fathers with newly created names, or grandsons, and so on. These names are *markers* that are likely to identify people with familial linkages.[4] If inheritance is important, so that these familial linkages connect people with relatively similar incomes, then the markers must make the same connections and, thus, be informative for income.

To make this clearer still, consider the evolution of the surname distribution versus the income distribution. They are (to this point) independent of one another. The frequency of a surname *cannot* be informative for income, in-and-of-itself. That is, what is *not* going on is that 'rich people

function $\mathcal{O}_t : \Omega \to [1, 2, \ldots, S]$ as that which ranks each surname according to its commonality so that, for each $k = 1, 2, \ldots, S$, $F_t\big(\mathcal{O}_t^{-1}(k)\big) \geq F_t\big(\mathcal{O}_t^{-1}(k+j)\big)$ for all $j > 0$ (ties are randomly allocated). $G_t(k)$, for $k = 1, 2, \ldots, S$, is then $G_t(k) \equiv F_t\big(\mathcal{O}_t^{-1}(k)\big)$. The long-run distribution is then defined as $G : [1, 2, \ldots, S] \to [0\,,\, 1]$ such that, for $k = 1, 2, \ldots, S$, $G(k) = \lim_{t\to\infty} E\big[F_t\big(\mathcal{O}_t^{-1}(k)\big)\big]$. Note that, since $F_t(s)$ is necessarily random for all t, we define $G(k)$ as the *expected* fraction of the population associated with the k^{th} most popular surname.

[4]For the same reason, surnames play an important role in the field of population genetics. The connection actually goes even farther. In our model, surnames are innocuous markers. They have no direct effect on income. Mitochondrial DNA, or the male Y-chromosome, is analogous. It does not code for any known protein and has no effect on the differential survival or reproductive chances of the individual receiving it. Nevertheless, it is a useful marker that allows researchers to uncover familial linkages.

have uncommon surnames.' What *is* going on is that low-frequency markers are indicative of familial linkages and high frequency markers are not. This is just as true for the rich as it is for the poor. [5]

We now turn to our main result, the behavior of the ICS. The ICS is a moment of the *joint* distribution of surnames and income. Even though the two are independent of one another, the ICS connects them and reveals information about the latter based on the markers inherent in the former.

Property 3 : ICS and the importance of inheritance

Under the conditions of Property 2 the ICS from equation (5) is approximately constant for all $t > k$*. Moreover, for any* $t > k$*, the ICS is monotonically increasing in the value of the inheritance parameter* ρ*.*

The proof is in Appendix A. The monotonicity result is analytic whereas the constancy result is shown via simulation.

Figure 2 plots the ICS against ρ for our baseline parameter values. Aside from confirming the monotonicity property of Property 3, what's quite striking is the level and the convexity. Relatively small values for the ICS are associated with moderately large values of ρ and only for very high values of ρ do we see ICS values above, say, 10%. Again, echoing comments made above, this is necessarily the case given that only the rare surnames can be informative.

To summarize, the results of this section are as follows. First, without mutations the number of surnames will tend to become small, with each name conveying very little information about familial linkages and, therefore, inheritance. Second, mutations provide a countervailing force, allowing many rare surnames to have informational content. Finally, the informational content of these rare surnames — the ICS from equation (5) — is, *ceteris paribus*, monotonically increasing in the magnitude of the inheritance parameter, ρ. This is what allows us to identify the magnitude of ρ from data on the joint distribution of surnames and economic outcomes. In our online appendix we demonstrate that these results (i) are robust to different parameter values, and (ii) are maintained in an extended model that features income-dependent fertility.

[5] In our online appendix we relax the independence assumption by allowing fertility rates and the mutation rate to depend on income. Our results do not change.

4 Data

We use data from two sources, the 2001 Spanish census and the 2004 Spanish telephone directory. From the census we have individual-level data from the Catalonian region of Spain on surnames, education and several other variables. From the telephone directory, obtained from Infobel, we have surname data. We describe the data and its uses below. First, however, it is important to understand how Spaniards name themselves and how this relates to our methodology.

4.1 Spanish Surnames

Spanish people have two surnames, 'first' and 'second.' The first is the first surname of their father and the second is the first surname of their mother. First surnames, therefore, are passed between generations in *exactly* the same manner as with the (traditional) Anglo-Saxon convention. Our methodology is based, primarily, on first surnames. It can be used in exactly the same way for males in Spanish societies as in many other Western societies.

This being said, the Spanish naming convention does offer additional information, information that we make use of. Unlike the Anglo convention, each male is connected with his mother. In addition, because females do not change their surnames upon getting married, each female is connected to her father. Note that this has no bearing on the evolution of the paternal lineage. The maternal surname vanishes after two generations.

We make use of this extra information in a number of ways. First, we use the second surname as a control for ethnicity, leaving the first as an indicator of the importance of familial linkages. Second, we use the combination of the two surnames to identify siblings and, thus, be more precise about familial linkages (Section 9). Third, we use the combination of the two names to identify the strength of 'assortative mating' and its importance for economic inheritance (Section 10).

4.2 Data Description

The census data for Catalonia covers the entire population of 6,343,110 individuals. For each individual we have their two surnames, some demographic characteristics (age, education, gender, marital status, place of birth, place of residence), as well as employment status, level of proficiency in the Catalan language and several housing characteristics (tenancy, size, inheritance, availability of a second house).[6] The census does not record information on wealth or income. We therefore

[6]We define place of birth differently for those born in Catalonia versus those born elsewhere in Spain. If Catalan-born, we use county dummies, otherwise we use Spanish province dummies. Catalan counties are administrative

use years of education as our measure of economic well-being, y from equation (6).

We eliminate individuals living in 'collective households' because the census has no educational information on them. We also eliminate individuals for whom the first or second surname is missing. This leaves us with 6,123,909 individuals. We eliminate females in order to be consistent with the literature on intergenerational mobility. We eliminate males who are less than 25 years of age so as to focus on individuals who have finished full-time education. We include only Spanish-born, Spanish citizens so as to mitigate the extent to which surnames are informative because they distinguish an immigrant who is likely to have relatively low education. Finally, we exclude individuals with a unique first surname because such names cannot, by definition, provide familial linkages with other individuals. This leaves us with our baseline population of 2,057,134 males.

We use data from the Spanish telephone directory (obtained from a commercial source) for the purpose of controlling for ethnicity (Section 5). There are roughly 14 million households in Spain and the directory contains surname information on roughly 11.4 million private, fixed telephone lines. Mobile phones, which are not included, obviously account for the majority of the difference. We have no reason to believe that the surname distribution differs across fixed versus mobile lines and are therefore confident that the absence of the surnames of mobile-only households does not affect our results.

4.3 The Surname Distribution

Figure 3(a) plots the commonality-ordered frequency distribution of the first surname (the empirical counterpart to G_t from Property 2 in Section 3.1). The distribution is very skewed. There exist a large number of low-frequency surnames and the few most frequent surnames represent a large percentage of the population. The 10 most popular names cover roughly 11 percent of the population.

Figure 3(b) plots the Lorenz curve for the Spanish and Catalan surname distributions, using both the telephone directory and census information for the latter. Table 1 reports the relevant statistics for the three distributions. Notice that the number of people per surname is larger in the whole of Spain than in Catalonia. This is probably because Catalonia is a net receiver of immigration and because the Catalan language has, historically, had less orthographic rigidity

units somewhat smaller than a typical U.S. county. Spanish provinces are somewhat larger than a typical French *departament*.

than the Spanish language. There are also more surnames in the census than telephone directory, probably because each household typically has just one telephone line.

5 Ethnicity

To this point we have emphasized the ability of surnames to reveal familial linkages. However they may also reveal ethnicity and this muddies the water. That is, sons inherit from their fathers (i) unobservable economic traits, (ii) observable ethnicity, and (iii) a surname. Ethnicity and surnames are often connected.[7] So are ethnicity and economic outcomes. Surnames might therefore be informative for economic outcomes simply because they are informative for ethnicity. Catalonia is a good example. It is well known that native Catalan speakers (roughly half the population) (i) have fairly distinctive surnames, and (ii) have high education and economic status relative to non-native-Catalan speakers.[8] This then begs the question, if surnames are informative in ethnically diverse societies, is this primarily due to ethnicity? Or do direct familial linkages also play a role?

We address this question by *controlling* for ethnicity and thereby examining mobility among ethnically-homogeneous partitions of the population. This allows us to answer the above question and ask how the answer has changed over time. It also represents a departure from much of the literature, so we must take care to interpret our measurements of the inheritance parameter, ρ from Equation (6), appropriately. The literature has typically estimated a parameter that *amalgamates* the effects of direct familial linkages with those of ethnicity. We prefer to separate the two so that (ideally) ρ measures only the direct familial effects. Controlling for ethnicity will underestimate ρ and therefore tend to overstate intergenerational mobility as it is typically defined in the literature. As such, we also report more comparable results in which we *do not* control for ethnicity.

[7]Surnames must eventually reveal ethnicity *even if there are no ethnic differences in the initial distribution of surnames*. This is because the surname birth/death process is independent across ethnic groups. Assume that there are only two ethnic groups, red and blue (r and b) with the red group being richer on average. A surname mutation among the r group will generate a new name which, until the name dies-off, will only be associated with the r ethnicity. A surname death among the b group will leave relatively more r ethnic-group individuals with this name, thus increasing informativeness about ethnicity. The independent birth/death process will lead the r and b surname distributions to drift apart over time. Eventually an individual's surname will necessarily be informative on whether his ancestors were of the r or b ethnicity. If ethnicity is related to other characteristics like income, then surnames will also be informative on these characteristics, even though they may not have to begin with. An example of paper that uses surnames to elicit ethnic information is Rubinstein and Brenner (2011).

[8]The reasons range from obvious to controversial. An obvious one is the initial condition. It is well known that immigrants into Catalonia have been considerably less wealthy and less educated than the native population. Controversial reasons include the linguistic advantage that native Catalan speakers have in the educational system (see Aspachs-Bracons, Clots-Figueras, Costa-Font, and Masella (2008)), and a variety of forms of discrimination that non-catalan speakers may be subject to.

Our methodology works as follows. First, by considering only Spanish-born, Spanish citizens we immediately eliminate non-Spanish ethnicities. This is because Spain (including Catalonia) saw very little foreign immigration prior to the 1990's. This leaves us with a Spanish population for which the primary ethnic trait is *regional origin*: the region of Spain from which one's family originates. We measure this with the following index of the *"Catalonianess"* of each particular surname s:

$$CatalanDegree(s) = \frac{\text{Number of telephones under surname } s \text{ in Catalonia}}{\text{Number of telephones under surname } s \text{ in Spain}},$$

which we interpret as an estimate of the fraction of people with surname s that reside in Catalonia. For example, since the 2001 Catalan population is about 16% of the total Spanish population, then, if surnames were uniformly distributed throughout Spain, this ratio would be roughly 0.16. The extent to which it's higher (lower) indicates a concentration of people with surname s residing inside (outside) of Catalonia.

We go one step further and interpret the $CatalanDegree(s)$ variable as a proxy for the extent to which a person with surname s has Catalan regional origin. In Appendix B we demonstrate that this variable is indeed a good proxy of ethnic regional origin. It has strong predictive power of both knowledge of the Catalan language and place of birth within Spain. It is not perfect, but it is the best that we can do with available data; and better than self-reported ethnic identity (which anyway is not available).

We use the $CatalanDegree(s)$ variable to control for ethnicity two different ways. First, in our ICS regressions we include on the RHS the $CatalanDegree(s)$ value associated with an individual's *second* surname (their maternal surname). A dummy variable for their *first* surname will be included for measurement of the ICS (Section 2, equation (5)). Using the first and second surnames in this manner is meant to mitigate multicollinearity. Second, in Section 6.1, we consider sub-populations of regionally-homogeneous groups. We calculate the geometric mean of the $CatalanDegree$ for both the first and second surname and then order the population according to the associated values. We then identify upper 50% quantile as a homogeneous group of individuals with Catalan regional origin.

6 Cross-Sectional Results

In this section we present results that are 'static' in nature. This means that we pool together individuals from different birth cohorts. We obtain cross-cohort *average* measures of the ICS. In Section 7 we use calibration to map them into measures of intergenerational mobility that are comparable to those in the literature. In Section 8, in contrast, we condition on birth cohort and obtain measures of how these things have changed over time.

6.1 The ICS in 2001

Table 2 reports a benchmark set of estimates of equations (3–4) and the associated ICS from equation (5). Column 1 begins by including only individual controls — dummy variables for age and place of birth. The adjusted R^2 is 0.2652. Column 2 adds our $CatalanDegree$ variable. The coefficient is positive and highly significant. It's also economically significant. The standard deviation of $CatalanDegree$ is about 0.3. Therefore, the estimate of 1.706 translates into an additional 0.5 years of education for a one-standard-deviation increase in a surname's 'Catalonianess.' The mean and standard deviation of education are 8.4 and 4.6, respectively. So Catalan regional origin is associated with higher educational attainment equal to about 10% of the overall dispersion.

Column 3 of Table 2 adds paternal surname dummies to the regression (recall that maternal surnames are used to define $CatalanDegree$). We note that (i) the surname dummies are jointly significant (given the large number of RHS variables involved this is not obvious in spite of the large population size), (ii) the coefficient of $CatalanDegree$ is smaller but remains economically meaningful, with a one-standard-deviation increase translating to 4 extra months of education, and (iii) the R^2 increases to 0.2980. Surnames are thus informative. Knowledge of the *particular surname* of an individual is informative for predicting their educational attainment.

Column 4 replaces the actual surname dummies with 'fake' dummies as in equation (4) of Section 2. The fake dummies are not jointly significant and their presence increases the R^2 very little. The estimate of the $CatalanDegree$ coefficient is largely unaffected. Our estimate of the ICS (controlling for ethnicity) from equation (5) is, therefore, 2.45%. Columns 5 and 6 are analogous to columns 3 and 4 except that the $CatalanDegree$ variable is omitted. Since surnames now capture both ethnicity and familial linkages, the ICS increases to 3.02%.

Tables 3(a) and 3(b) repeat the exercise of table 2, but restricting the population to those born in Catalonia (table 3(a), immigrants are not included, even if their children are) and to those born

in Catalonia before 1950 (table 3(b)). The results are very similar to our benchmark in Table 2, the only exceptions being that the R^2s are smaller and the ICS is slightly higher. Both are to be expected given that the population is more ethnically homogeneous.

One concern is that, in spite of our use of $CatalanDegree$, our results are dominated by ethnic and not familial linkages. To examine this, Table 3(c) restricts the sample to the 50% of the population who have surnames with the highest $CatalanDegree$. The qualitative nature of our results is unaffected, with the ICS being just slightly higher. Note that, encouragingly, the ICS is basically unaffected by the inclusion of the $CatalanDegree$ variable.

To summarize, our results show that surnames are informative for educational attainment. Part of this is because surnames are informative for Spanish regional origin. But an important part remains, even after controlling for regional origin in a variety of different ways. Our interpretation — that elucidated by our model — is that the surname partition is informative about familial linkages. We now present two additional pieces of evidence that support this interpretation.

6.2 Rare Surnames are More Informative

Our model predicts that, if inheritance is important, surnames will be informative because the partition of *rare* surnames should group together people with familial linkages. This then implies that the ICS should *increase* as we exclude common names. Checking this in the data provides a valuable check on our interpretation of our results.

Table 4 repeats the exercise of Section 6.1 but includes only the 50% of the population with the least-frequent surnames. As our model predicts, the ICS increases. It does by a factor of roughly 1.5 (with and without ethnicity controls) relative to Table 2. Figure 4 provides additional evidence. It is based on a series of regressions, analogous to those in Table 2, that sequentially include people with more and more common surnames, as we move from left to right on the horizontal axis. The left vertical axis reports the ICS — the downward-sloping line — and the right axis reports the average number of individuals per surname in each sub-population. Figure 4 provides strong evidence in favor of our model's interpretation of the ICS. Moving from least frequent names to the most frequent, the ICS monotonically falls by a factor of five. This suggests that our controls for ethnicity are working and that our findings are driven by the informativeness of surnames for familial linkages.[9]

[9]It is important to understand what Figure 4 says and what it does not say. It says that as average surname

6.3 Invented Catalonias

Our results should not be sensitive to any random (but sufficiently large) partitioning of the surnames set. One such partition is simply based on the alphabet. If, for example, we randomly assign letters of the alphabet to two groups, "first half" and "second half," then the ICS should be unaffected. This is what we find in Appendix C. This suggests that our findings are structural and that they depend on deeply rooted social and economic mechanisms.

7 Calibration

We calibrate the model of Section 3 using the results of Section 6. The main task is to obtain a value of the inheritance parameter, ρ, (as well as the parameters that determine the birth-death process of surnames) that results in our model's ICS matching the value of 3.02% from Table 2 of Section 6.1 for a surname distribution that simulates the one that we observe in Catalonia.

We choose our model's parameter values as follows. In the model of Section 3 we look for the combination of three parameters (ρ, mutation rate and family size) in order to minimize the distance with three moments of the data: (i) the ICS of years of education without controlling for ethnicity, (ii) the Gini Index of the surname distribution and (iii) the number of people per surname. The remaining parameter of the model (V_ϵ) is always chosen so that the unconditional standard deviation of education equals the one observed in the data (4.655).

The model does not fit perfectly to the data. The Gini Index of the surname distribution is larger in the data than the maximum that can be obtained in any of the simulated economies. We proceed by minimizing a weighted distance of the moments of the artificial economy with the ones of the data. The exercise (reported in Table 5) results in a value of ρ of 0.57. This value of ρ is very robust to changes in the specification and in the weights of the function we minimize.[10]

This result is of interest along several dimensions. First, it substantiates the point made at the end of Section 2; small 'incremental R^2' units for the ICS map into large 'inheritance units' for ρ. This is the nature of our method and it's a natural consequence of extracting information from

frequency decreases, the informativeness of the *individual surnames* increases. It does not say anything about the informativeness of the *individual surname frequencies*. The latter would ask, for example, 'is someone with a rare surname likely to be highly educated?' This is *not* what we are asking here. It *does*, however, get addressed in our online appendix. We find that (i) surname frequencies *are* mildly informative, and that (ii) the important properties of the ICS are nevertheless unaffected.

[10] An interactive Gauss program for determining the calibration is available from the authors upon request. Moreover in our online appendix we provide with a detailed explanation of the calibration process. We also calibrate an extension of the model allowing for differences in fertility which results in approximately the same value of ρ.

a highly skewed surname distribution. Second, $\rho = 0.57$ is somehow larger but in line with the few estimates available for Spain.[11] Third, this number is similar in magnitude to the analogous estimates — based on very different data and methodology for other countries.[12]

8 Dynamic, Cohort-Based Results

Our analysis to this point has treated the entire 2001 Catalan census as a single cross-section. The cross-section, of course, consists of individuals of different ages. We have dealt with education-related age effects using dummy variables. But we have not allowed the ICS to vary with age. The above estimates are age-averaged measures of the ICS. They are 'static' in the sense that they are incapable of saying anything about how the ICS and intergenerational mobility may have changed over time. As mentioned before, little is known about the time evolution of mobility.[13]

We turn now to a dynamic analysis. We partition the cross-section into birth cohorts and ask if the ICS varies across subpopulations of different ages. If it does, our model suggests that intergenerational mobility may have changed over time. The reasoning is as follows. Suppose that mobility has decreased, so that the value of ρ connecting the current young generation to their parents has increased. Our model's prediction is that, if the surname distribution of the parents is highly skewed (which it is), then the surnames of the young should be more informative than those of the old. The change in mobility should generate a change in the ICS. Why? For the same reasons as above. A higher value of ρ implies that that familial linkages will be more informative for economic status. A skewed surname distribution of parents means that the surname distribution of children will be informative for familial linkages.

There are, of course, alternative interpretations for why the ICS may have changed over time. One possibility is a change in the birth/death process for surnames. We discount this possibility

[11] Kalkbrenner and Villanueva (2007) estimate educational mobility for Spain from the "Encuesta de Conciencia y Clase de Biografia" for 1990-91 and obtain an estimate between 0.42 and 0.52 among fathers and sons. The other estimation of which we are aware for Spain (Pascual (2009)) is for income mobility using the European Community Household Panel and reports an elasticity of 0.3 among fathers and sons.

[12] Solon (1992) addressed a number of the aforementioned issues around the estimation of mobility using panel data and argued for a mobility estimate of 0.4 or higher. A number of subsequent studies have reached similar conclusions based on different data. A value between 0.4 and 0.6 seems somewhat of a consensus, at least based on U.S. data for the latter third of the 20th century.

[13] The short time dimensions of the panel datasets make it very difficult to assess the dynamics in mobility. Moreover, Lee and Solon (2009) and Hertz (2007) attribute a fairly divergent body of existing results to small-sample bias in addition to the aforementioned age bias and sample attrition problems (*c.f.*, Mayer and Lopoo (2005) and Fertig (2004)). Taking this into account leaves the authors inconclusive about trends in intergenerational mobility. One exception is Blanden, Goodman, Gregg, and Machin (2004), who argue for a decrease in mobility in the U.K. between two cohorts of people born in 1958 and 1970, respectively.

for one simple reason. Such a change must necessarily affect the ICS via its effect on the surname distribution. This effect will occur very *slowly*. Its immediate effect on the ICS, therefore, must be very small. A change in ρ, in contrast, will have an *immediate* impact on the ICS because the channel through which it works is not the slow-moving surname distribution. Our empirical robustness checks (in Section 9) and our assortative mating exercise (in Section 10.2) serve to bolster this point and provide reassurance that we are capturing changes in ρ.

Figure 5 reports our results. We run the same set of regressions as those in Table 2 for a rolling sequence of overlapping 25-year age cohorts. Figure 5 reports the evolution, from oldest cohorts to youngest cohorts, of the ICS and the point estimate of the parameter on our $CatalanDegree$ variable. What we see is striking. Figure 5(a) shows that the ICS is substantially higher for younger cohorts. Figure 5(b) shows that regional origin has become more important for determining educational outcomes. Figure 6 reports the evolution of the ICS when not controlling for ethnicity. As expected, the ICS is higher when it amalgamates familial and ethnic effects (Section 5).[14] Finally, Figure 7 conducts the same sort of robustness checks that were undertaken for the single cross-section in Section 6. The temporal behavior of the ICS is qualitatively the same and quantitatively larger when we restrict the population to those with high $CatalanDegree$ surnames (a much more homogeneous group in the ethnic dimension) and those low surname frequencies.

We interpret these results as indicating that intergenerational mobility has decreased in Catalonia, both because of an increased importance of ethnicity and an increased importance of familial linkages for determining educational outcomes. In Section 9 we substantiate this interpretation further by examining subpopulations of siblings. In Section 10 we offer and substantiate a potential explanation: increased assortative mating. Before doing so we offer some interpretative discussion of these findings as they relate to important historical trends in 20th-century Catalonia.

8.1 Public Education

Decreased mobility in Catalonia is particularly striking when held against the backdrop of the large, secular changes that occurred in publicly-provided education in Catalonia and Spain during the mid 20th century. Others have also found a similar result (*c.f.*, Checchi, Ichino, and Rustichini (1999), Grawe (2010) and Parman (2011)). In Spain, prior to the 1950's access to education was very limited. This was a consequence of both the general level of wealth and income as well as

[14] Our online appendix provides regression details and additional results for a more coarse set of cohorts, those born before/after 1950.

the lack of investment in public education. Starting in the late 1950's things began to change. Both the economy and the level of public education began to grow, particularly from 1975 onward. This shows up clearly in our data. Figure 8 plots the mean and standard deviation of the years of education for the same overlapping sequence of 25-year age-cohorts that we used above. It demonstrates that the average years of education of the oldest individuals in 2001 is less than half that of the youngest. Variability around the average, in contrast, is more stable.

How is 'more education for all' associated with intergenerational mobility in educational attainment? Does an increase in publicly-provided education decrease the importance of how educated one's parents are? While the intuitive answer might be yes, our results show that intuition does not fit the facts. Such intuition confuses aggregate growth with cross-sectional mobility. Mobility is a *relative* concept. Aggregate growth is not.

8.2 Increase in the Importance of Ethnic Background

Our point estimates of the coefficients on the $CatalanDegree$ variable have increased, indicating an increase in the importance of ethnicity (Figure 5(b)). Note that this cannot be a direct consequence of the migration process. In our regressions we include not only the children of immigrants, but also the immigrants themselves. Thus, our result indicates that $CatalanDegree$ is more important for measuring the education of the second than of the first generation of immigrants. It is not the case that it increases because there are more immigrants. Note also that this does not mean that low $CatalanDegree$ individuals have obtained less education, but that their difference vis-à-vis high $CatalanDegree$ individuals has increased. The increase in educational attainment has been large for both ethnic groups, but has affected Catalan speakers more than non-Catalan speakers.

In Catalonia there are two main linguistic communities, Catalan and Castillian (Spanish), each representing roughly half the population. Catalan speakers have enjoyed substantially larger incomes and larger levels of educational attainment during the entire period of our study (this is true for both those born before and after 1950). Nevertheless before the late 1970's there did not exist any formal linguistic advantage toward Catalan speakers. The language of government, commerce and education was overwhelmingly Spanish. However, beginning in the late 1970's the increasing political power of Catalan nationalism has translated into a series of drastic legal and administrative reforms that have turned upside down the relative importance of both languages in

society while changing only marginally its overall language composition.[15] For example, since the beginning of the 1980's all education is provided *exclusively* in Catalan in all public and practically all private schools. Catalan is now the sole language of the regional and municipal governments, and proficiency in Catalan has been the key requirement for working in public administration since the beginning of the 1980's. Further legal change has made Catalan an important (albeit perhaps not the main) business language.

Governmental and institutional changes in the use of Catalan are, at best, a partial explanation for what we find. A deep understanding of the increase in the value of ethnicity is beyond the scope of this paper. Note, however, that in Section 10.2 we do dig a little deeper and show that *assortative mating* seems to have increased in Catalonia along ethnic lines. This makes ethnicity more inheritable and serves to magnify the effects of any institutional changes on the educational outcomes of the offspring of those who assortatively mate.

9 Analysis of Siblings

An established alternative to father-son correlations is sibling correlations (Solon, Corcoran, Roger, and Deborah (1991)). The reasoning is simple. If economic inheritance is important then the outcomes of siblings should be correlated because they share parents and, thus, they share the same inherited economic traits. In this section we make use of this logic to provide an important robustness check on our results. In doing so, we also draw links to an established branch of the literature. Using the peculiarities of the Spanish (second) surname we can approximate siblings with a high degree of accuracy and provide an alternative measure of mobility. We then compare this qualitatively different measure to the ICS.

We identify siblings in the following manner. Recall that all Spaniards have two surnames, one from their father and one from their mother. Thus, all siblings (irrespective of gender and marital status) share not one but two surnames, as well as their ordering. This allows us to construct a partition of the population that groups together individuals who have a high likelihood of being siblings.

[15] See Miley (2004) for a study of the politics of nationalism and language in Catalonia. The increasing power of Catalan nationalism might be explained (i) by the larger levels of income and education of the Catalan speaking community and (ii) because Spanish electoral law has allowed Catalan nationalism to operate as a third party in Spanish politics, allowing it to obtain high leverage from its successive alliances with either left or right leaning governments. See also Aspachs-Bracons, Clots-Figueras, Costa-Font, and Masella (2008) for a study of the effects of linguistic legislation on the educational system on identity.

More specifically, define the "complete–surname" for an individual to be their two surnames, in order. That is, if a person's father and mother are named Fernández and Caballé, respectively, then their complete–surname is "Fernández Caballé". This is distinct from both "Caballé Fernández" and "Fernández Vila". Next, group each person together with those who share their complete–surname, and eliminate anyone whose complete–surname is unique in the population. This partition will be very similar to the actual sibling partition (which we cannot observe), with the similarity increasing in the *rarity* of the surnames. The reason is that it's very unlikely for two males who share the same rare surname to marry two females who share the same rare surname, thus generating children *who are not siblings* with the same complete–surname. What is much more likely is that two individuals with the same *rare* complete–surname are in fact siblings. We therefore focus on matches of rare, complete-surames.[16]

We proceed as follows. We form a partition of the subpopulation of those who share their complete-surname with either one or two other individuals. As above, we run two regressions, one with legitimate complete–surname dummy variables and one with fake dummy variables (defined as in Section 2). We define the *Informational Content of Siblings* (ICSIB) as the difference between the R^2's from these regressions. Our results are reported in Table 6(a). As expected the ICSIB is much higher than the ICS from previous sections. Table 6(a) cannot control for ethnicity because our *CatalanDegree* variable and the complete–surname dummy are based on a common surname. Table 6(b) therefore reports on a subsample containing only the 50% most Catalan surnames, resulting in a more ethnically homogeneous population. Our results do not change.

Figure 9 and 10 report results that serve as a powerful robustness check on our method and its findings. Figure 9 shows the evolution of the ICSIB over time for those who share their complete–surname only one other person, and for those who share it with at most two other people. Figure 10 shows the same thing for the subpopulation with the 50% most Catalan surnames. In both cases we see a marked increase over time in the ICSIB, the same patter as we observed in the ICS of Section 8. The two measures are conceptually and mechanically quite different from one another. Yet our model and the existing literature suggest that they are both legitimate measures of changing mobility. The fact that they tell the same story provides important reassurance to our interpretation of the ICS.

[16]In a previous working paper we provided results that varied across 'rarity' in this context. We showed that the more people there are who share a complete surname — thus making for a higher likelihood of non-siblings appearing in the set — the lower is the ICS.

10 Assortative Mating

In this section we develop and test one possible explanation for our finding that mobility has decreased over time: an increase in *assortative mating*.

Assortative mating refers to the tendency of people with similar characteristics to marry each other.[17] At first blush, it may seem intuitive that assortative mating can give rise to the ICS because, ostensibly, it can generate "organization" in the distribution of surnames. If, for instance, today's rich and poor have distinct surnames, and if the rich marry the rich and the poor marry the poor, then one might think that the rich and poor surnames will remain distinct among future generations, thus generating informativeness. One might apply a similar argument to ethnically-motivated assortative mating. In either case, this intuition is deeply misleading. This is because the degree of assortative mating does not have any *direct* effect on the *marginal* distribution of surnames in the population. The reason is simple. Surnames are passed along the male lineage. For surname determination, it does not matter *why* one's father married one's mother, all that matters is one's father's name. It is as if females had no surnames.[18]

What assortative mating *does* matter for is the *joint* distribution of surnames and characteristics. This is because, if inheritance occurs along both the maternal *and* the paternal lineage, more assortative mating amounts to the increased prevalence of inheritance. In the language of our model, more assortative mating increases the father-son inheritance parameter, ρ, and, via this mechanism, it increases the ICS. We now enrich our model to make this mechanism precise. The model is richer in that it articulates the mapping between assortative mating and the inheritance parameter that we have estimated above. The model does not, however, have any bearing on the estimates themselves. This is because, in our enriched model, *if we take ρ as given*, the process that generates the joint distribution of surnames and income is *identical* to that of Section 3.

10.1 A Model of How Assortative Mating Affects Inheritance

We now treat Equation (6), which correlates the income of *fathers* and *sons*, as endogenous. What is exogenous is (i) the manner in which the income of *sons* and *daughters* depends on the income of both their father *and their mother*, and (ii) the manner in which boys and girls sort at the time

[17] The existing literature on mobility that incorporates assortative mating includes Lam and Schoeni (1993), Chadwick and Solon (2002), Ermisch, Francesconi, and Siedler (2006) and Holmlund (2006). There is also a rich literature in macroeconomics that focuses on assortative mating and inequality (*e.g.*, Fernández and Rogerson (2001), Fernández, Guner, and Knowles (2005) or de la Croix and Doepke (2003)).

[18] We are thankful to Melvin Coles for this insight.

of mating. There is a *continuum* of males and females who form households and bear offspring. Expanding on the notation, y^m_{ist} and y^f_{ist} denote the incomes of male and female children who inhabit household i with paternal surname s at date t. This household was formed at date $t-1$. The children's incomes arise as

$$y^m_{ist} = rz_{ip,t-1} + e^m_{ist} \quad ; \quad y^f_{ist} = rz_{ip,t-1} + e^f_{ist} \quad , \tag{7}$$

where z is family income, which we assume to be the average income of the father *and* mother's income: $z_{ip,t-1} \equiv (y^m_{ip,t-1} + y^f_{ip,t-1})/2$. The e innovations are *i.i.d.* $\mathrm{N}(0, V_e)$ and $r \in (0,1)$ is a *household* inheritance parameter. The distribution of z is endogenous. We prove its existence below. The parameters r and V_e are exogenous and determine the process of income transmission *given* the income of the two parents.

Mating determines the distribution of *family* incomes for the subsequent generation, given the distribution of incomes of sons and daughters of the current one. The incomes of the current generation's male children become that of the next generation's fathers: $y^m_{ipt} = y^m_{ist}$. Each father forms a household with a female who becomes a mother. The income of the mother is described by a *mating technology*, a function $f(y^m, u)$ that combines each father's income, y^m_{ipt}, with a *mating shock*, u_{ipt}, to assign to each father a spousal income, y^f_{ipt}, such that the distribution implied by $f(y^m, u)$ coincides with that implied by the inheritance process (7) for the population of female children at date t. The function that we use is

$$y^f_{ipt} = \lambda y^m_{ipt} + u_{ipt} \quad ; \quad u_{ipt} \sim N\left(0, V_u\right) \quad , \tag{8}$$

where $\lambda \in (0,1)$ — the correlation between spousal incomes — is the degree of assortative mating. Note that Equation (8) is silent on the particular assignment mechanism that 'mates' the distributions of y^m_{ist} and y^f_{ist} from Equation (7). For our purposes, it is sufficient to simply form a set of ordered pairs, $(y^m_{ist},\, y^f_{ist})$, that satisfy two properties: (i) they capture the notion of assortative mating that we are interested in, and (ii) they are consistent with the distributions implied by the inheritance processes (7). Examples of more fully-articulated assignment mechanisms are in Becker (1973), Gavilán (2012), Kremer and Maskin (1995), Marimon and Zilibotti (1999) and Shimer and Smith (2000).

In Appendix A we show that Equations (7) and (8) imply a stationary distribution for family

income, z. Inspection of the inheritance processes, Equation (7), then implies that, since $r < 1$, the stationary distributions of male and female income must be the same,

$$y^m_{ist} \sim N(0, V_y) \; ; \quad y^f_{ist} \sim N(0, V_y) \quad , \tag{9}$$

where V_y is a unique function of the model's structural parameters, r, V_e and λ. This function is characterized in Appendix A. We assume that the variance of the mating shock is $V_u = (1-\lambda^2)V_y$. This guarantees that the distribution of brides has the same cross-sectional distribution as that of female income.

Note that the inheritance parameter r from Equation (7) relates male children's income to the income of their parents' *household*. The model of Section 3 and most of our empirical work, in contrast, refer to the correlation between children and their *father*. This is because surnames are passed along only the male lineage. Therefore, in order to understand how assortative mating affects the ICS we must describe how the parameters r, V_e and λ are manifest in both the variance of the income V_y and the parameter ρ from the following expression:

$$y^m_{ist} = \rho y^m_{ip,t-1} + w^m_{ist} \quad , \tag{10}$$

where the variance of w is denoted V_w. This equation links the income of sons to their fathers, a relationship that depends on both the mating process, (8), and the household-level inheritance process (7).

Note also that, for issues of intergenerational mobility, the appropriate measure of inheritance is ρ and not r. This is because ρ associates comparable variables — the incomes of children with their father — whereas r from Equation (7) does not. The latter associates the income of one individual with the consolidated income of their childhood household, something that arises from the noisy lottery of mating.[19]

In Appendix A we prove the following property:

Property 4 *There exists a unique stationary distribution for y^m_{ist} and y^f_{ist} that is characterized by*

$$\rho = \frac{r(1+\lambda)}{2} \; ; \qquad V_w = V_e \left(1 + \frac{r^2(1-\lambda)}{4\lambda} \right) \; ; \qquad V_y = \frac{V_e}{\lambda(1+\lambda)}$$

[19] One could, alternatively, use an analogous parameter that associates the consolidated income of each household with the consolidated income of that household's children's households. Indeed, a number of existing studies on mobility and assortative mating do just this. We choose to focus on ρ from Equation (10) because (i) it is perfectly coherent, (ii) it is the parameter that is estimated in the most of the existing literature on mobility, and (iii) it is tightly linked to the process of surname diffusion.

A larger degree of assortative mating — as measured by a larger value for λ, the correlation of spousal income — thus translates into a larger value of ρ. Stronger assortative mating implies less intergenerational mobility in the population of fathers and sons. This is true even if the correlation between the income of sons and the joint income of their parents, r, is held constant. The intuition is straightforward. More assortative mating implies that the father's income is more informative for the income of the mother. Both father and mother contribute to the characteristics of their son. Thus, the more the income of the father explains the income of the mother, the more it must explain the income of his son. Stronger assortative mating translates into lower intergenerational mobility.

The same intuition applies to any other inheritable trait, like ethnicity. Females, of course, play an important role in determining the ethnicity of a household's children. Keeping in mind that surnames capture ethnicity *only insofar* as it is transmitted across the male lineage, it becomes clear that assortative mating is pivotal. It is the only way with which a mother's ethnicity can be correlated with her children's surname. Consider, for example, Judaism in which (ignoring conversion) ethnicity is *solely* passed along the maternal line. Absent assortative mating — *e.g.*, if Jewish women marry men randomly drawn from the entire population of males — surnames must eventually become uncorrelated with the Jewishness of their holders. On the other hand, if Jewish women marry only Jewish men, then the surnames of Jews will become increasingly distinct from those of gentiles, owing both to the initial distribution of Jewish male surnames and to the surname birth/death process described previously. This mechanism applies to virtually any other ethnicity-related characteristic. Since females are in almost all cases fundamental for the inheritance of ethnic characteristics, assortative mating and the degree of ethnic information contained in surnames go hand-in-hand.

To summarize, surnames are passed exclusively along the male line. They do not provide any *direct* information about the mother. Any information that is *indirectly* associated with the mother must arise because the characteristics of the father are correlated with those of the mother. This is the mechanism through which assortative mating can affect the ICS. In the language of our model, the ICS depends *only* on the correlation and conditional variance of the incomes of fathers and sons: the parameters ρ and V_w, respectively. But assortative mating affects ρ and, therefore, it affects the ICS. This lends valuable interpretation to our empirical findings in the next subsection.

10.2 Assortative Mating: Evidence

We have seen that surnames contain information on two characteristics: ethnicity and educational attainment. This, combined with the Spanish naming convention, allows us to obtain measurements of the level and change in ethnic/educational assortative mating in Catalonia. As we have seen, an increase in the degree of assortative mating translates into an increase in the prevalence of inheritance, and of the ICS.

Our identification strategy is best illustrated with an example. The surname Casals is associated with a high value of our *CatalanDegree* variable. The same applies to the surname Pujol. A person whose complete–surname is "Casals Pujol" — a person whose father is Casals and mother is Pujol — is therefore almost certainly a person with two parents of Catalan regional origin. Ethnic assortative mating, then, can be measured by the incidence of such complete surnames relative to those that are more ethnically-heterogeneous. The measurement is simple correlation between the ethnicity index of each person's first and second surname.

Note that this measurement applies to each individual's *parents*, not to each individual's spouse. That is, if we find evidence of increased assortative mating among the 25-30 year-old cohort in the 2001 Catalan census, this means (very roughly) that the 50-55 old cohort exhibited more assortative mating than those one generation older.

The data are constructed by first associating to each surname two characteristics: the average level of education and the average value of *CatalanDegree*, where the average is taken across all individuals with that particular surname. We then run two sets of regressions, one for each characteristic. The LHS variable is each individual's first surname's characteristic and the RHS variables are the set of controls used above along with their second surname's characteristic.

Figure 11 displays the results. We plot the value of the parameters for the regression of education (Figure 11(a)) and for the regression on *CatalanDegree* (Figure 11(b)) for the moving window of cohorts described above.[20] Assortative mating in both characteristics is clearly increasing, and education has a timing that resembles the timing of the increase in the ICS.

In order to make sure that our results are not driven by ethnicity we run the same regressions on ethnically homogeneous populations (Figure 12(a)) and on populations with very infrequent surnames (Figure 12(b)). The results are qualitatively identical.

To summarize, we have found evidence that intergenerational mobility in educational attain-

[20] Our online appendix reports the same exercise splitting the population into those born before/after 1950.

ment has decreased in Catalonia in the 20th century. One possible explanation is that assortative mating has increased. Surname data is consistent with this explanation, suggesting an increase in the likelihood that people mate with others of similar educational levels and ethnic backgrounds.[21]

11 Conclusions

Our paper makes two contributions, one methodological and one applied. Methodologically, we develop a framework that shows how an untapped data source can shed light on a question that requires much data, but for which relatively little data exists. We show that a single cross-sectional census can reveal much about both the level and the change in intergenerational mobility. The key data objects are surnames, *markers* that provide intergenerational links where more explicit links are unavailable. Surnames define a *partition* of a population. Elements of this partition associated with *rare* surnames will be correlated with the partition that groups people according to familial linkages. A particular moment of these partitions — that which we label the *Informational Content of Surnames* (ICS) — connects the familial linkages with familial economic status and thus provides information on intergenerational mobility. This method yields measures of the degree of mobility at a point in time as well as its evolution across time. [22]

Our method would be of limited practical value in the presence of multi-country, intergenerational panel data. However the existence of such data is quite limited. The practical relevance of our method, therefore, depends on how much data we *do* have on the joint distribution of surnames and economic outcomes. Here, there is reason to be optimistic. Most countries compile censuses containing such data. We've shown that one can learn much from *one* census. Multiple census,

[21] Some existing work on increased assortative mating attributes it to an increased level of education among females. As above, one needs to be careful not to confuse this story with the effect of an increase in average educational attainment. Suppose, for instance, that the primary driver of assortative mating is wealth. Suppose also that there has been no change in the tendency for people to assortatively mate. If the daughters of the rich experience an increase in education that is larger than the daughters of the poor — something that is very plausible — then one might mistakenly conclude that assortative mating has increased, in the educational dimension, although in reality it has not. Our methodology does not suffer from this possible bias because our measures do not refer to the individual woman, *but to her family.* Education and ethnicity are imputed by the surname, not measured at an individual level.

[22] Several studies estimate mobility for a given country (Lillard and Kilburn (1995), Dearden, Machin, and Reed (1997), Wiegand (1997), Osterberg (2000), Osterbacka (2001), Hertz (2001), Dunn (2007), Ferreira and Veloso (2006), Leigh (2007), Ng (2007)). And a growing body of work has attempts to compare mobility across countries: several studies compare the USA to other countries (Björklund and Jäntti (1997), Couch and Dunn (1997), Checchi, Ichino, and Rustichini (1999), Björklund et al. (2002), Grawe (2004)); and Comi (2003) uses the European Community Household Panel to obtain estimates for 12 EU countries. Nevertheless, it is important to keep in mind the problems that plague cross-country analysis (Solon (2002)). Our method may represent promise in this context. Our model can be calibrated to specifically incorporate some of the cross-country heterogeneity that forms some of the basis of Solon's critique. We leave this to future research.

both within and across countries, can obviously yield much more. *Comparability*, over time and across countries, can be handled. The US, for example, has a different surname distribution than Spain. The essence of our method — the idea that rare surnames connect people with familial linkages — is nevertheless unaffected.

Our practical contribution is to use our methodology to ask how and why intergenerational mobility has *changed* over time. We study Catalonia, a large region of Spain. Using the 2001 census we show that the explanatory power of surnames — the ICS — has increased. Part of this is due to the increased explanatory power of ethnicity. But there is more going on. There is a component of the ICS that is *unrelated* to ethnicity and the impact of this component has also increased. This is true among very ethnically-homogeneous individuals, among siblings, and among people with infrequent surnames. Our model, alongside an extensive set of controls and robustness checks, associates this increase in the ICS with a decrease in intergenerational mobility. Our model and data also offer one possible explanation. Assortative mating along the ethnic dimension appears to have increased in tandem with the decrease in mobility.

To wrap-up, we offer some historical context. In Spain and Catalonia, the different generations of the 20th century witnessed large-scale *increases* in both the level of publicly-provided education and the level of educational attainment. Nevertheless, we've found that educational mobility has *decreased*. That is, the importance of family-specific characteristics for educational outcomes has *increased*. Is their a logical contradiction here? If one looks around and sees that almost everyone's educational attainment exceeds that of their parents, does this mean that the importance of inheritance and familial linkages must have diminished? The answer is no. Such logic confuses aggregate growth — an increase in the *mean* of the distribution — with mobility, which is all about movement *within* the distribution. It is at the heart of the common misperception that to do better than one's parents means to have beaten the odds and done better than expected. This can generate an upward bias in our perception of intergenerational mobility in growing economies. It is an illusion. It is just growth. Mobility works along its own path. It is defined only in relative terms. To measure mobility, it is not enough to compare my welfare with that of my parents. I must also consider the children of other parents, parents that were both richer and poorer than my own. Today's generation may well live better than yesterday's, while at the same owing a greater thanks to their parents for their place in the cross-sectional distribution.

References

Aaronson, D. and B. Mazumder (2008). Intergenerational economic mobility in the U.S.: 1940 to 2000. *Journal of Human Resources 43*(1), 139–172.

Angelucci, M., G. De Giorgi, M. Rangel, and I. Rasul (2010). Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics 94*(3-4), 197–221.

Aspachs-Bracons, O., I. Clots-Figueras, J. Costa-Font, and P. Masella (2008). Compulsory language educational policies and identity formation. *Journal of the European Economic Association 6*(2-3), 434–444.

Bagüés, M. F. (2005). ¿Qué determina el éxito en unas oposiciones? Fedea, Documento de Trabajo 2005-01.

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political Economy 81*, 813–846.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review 94*(4), 991–1013.

Björklund, A., T. Eriksson, M. Jäntti, O. Raaum, and E. Österbacka (2002). Brother correlations in earnings in Denmark, Finland, Norway, and Sweden compared to the United States. *Journal of Population Economics 15*(4), 757–772.

Björklund, A. and M. Jäntti (1997). Intergenerational income mobility in Sweden compared to the United States. *American Economic Review 87*(5), 1009–1018.

Black, S. E. and P. J. Devereux (2011). *Recent Developments in Intergenerational Mobility.* in Orley C. Ashenfelter and David Card (eds.), Handbook of Labor Economics, Volume 4B, Amsterdam: North-Holland, pp. 1487-1541.

Blanden, J., A. Goodman, P. Gregg, and S. Machin (2004). *Changes in Intergenerational Mobility in Britain.* in Miles Corak (ed.), Generational Income Inequality, Cambridge University Press, pp.122-146.

Cabré, A. (2004). La aportación de los 'otros' catalanes. El País (Edición Barcelona), 06/09/04.

Chadwick, L. and G. Solon (2002). Intergenerational income mobility among daughters. *American Economic Review 92*(1), 335– 44.

Checchi, D., A. Ichino, and A. Rustichini (1999). More equal but less mobile? education financing and intergenerational mobility in Italy and in the U.S. *Journal of Public Economics 74*(3), 351.93.

Clark, G. (2013). What is the true rate of social mobility? evidence from the information content of surnames. mimeo.

Collado, M. D., I. Ortuño-Ortín, and A. Romeo (2012a). Intergenerational linkages in consumption patterns and the geographical distribution of surnames. *Regional Science and Urban Economics 42*, 341–350.

Collado, M. D., I. Ortuño-Ortín, and A. Romeo (2012b). Long-run intergenerational social mobility and the distribution of surnames. mimeo.

Comi, S. (2003). Intergenerational mobility in Europe: evidence from ECHP. mimeo.

Couch, K. A. and T. A. Dunn (1997). Intergenerational correlations in labor market status: A comparison of the United States and Germany. *Journal of Human Resources 32*(1), 210–32.

Dahan, M. and A. Gaviria (2001). Sibling correlations and intergenerational mobility in Latin America. *Economic Development and Cultural Change, University of Chicago Press 49*(3), 537–54.

de la Croix, D. and M. Doepke (2003). Inequality and growth: Why differential fertility matters. *American Economic Review 93(4)*, 1091–1113.

Dearden, L., S. Machin, and H. Reed (1997). Intergenerational mobility in Britain. *Economic Journal 107*, 47–64.

Duncan, O. D., D. Featherman, and B. Duncan (1972). *Sociological Background and Achievement*. New York: Seminar Press.

Dunn, C. (2007). The intergenerational transmission of lifetime earnings: Evidence from Brazil. *The B.E. Journal of Economic Analysis & Policy 7,*(Iss. 2 (Contributions)), Article 2.

Ermisch, J., M. Francesconi, and T. Siedler (2006). Intergenerational mobility and marital sorting. *Economic Journal 116*, 659–679.

Fernández, R., N. Guner, and J. A. Knowles (2005). Love and money: A theoretical and empirical analysis of household sorting and inequality. *Quarterly Journal of Economics 120 (1)*, 273–344.

Fernández, R. and R. Rogerson (2001). Sorting and long-run inequality. *Quarterly Journal of Economics 116 (4)*, 1305–1341.

Ferreira, S. G. and F. A. Veloso (2006). Intergenerational mobility of wages in Brazil. *Brazilian Review of Econometrics 26*(2), 181–211.

Fertig, A. R. (2004). Trends in intergenerational earnings mobility in the U.S. *Journal of Income Distribution 12*, 108–130.

Fryer, R. and S. Levitt (2004). The causes and consequences of distinctively black names. *Quarterly Journal of Economics 119*(3), 767–805.

Gavilán, A. (2012). Wage inequality, segregation by skill and the price of capital in an assignment model. *European Economic Review, forthcoming 56 (1)*, 116–137.

Grawe, N. D. (2004). *Intergenerational mobility for whom? The experience of high- and low-earnings son in international perspective.* in Miles Corak (ed.), Generational Income Inequality, Cambridge University Press, pp.58-89.

Grawe, N. D. (2010). Primary and secondary school quality and intergenerational earnings mobility. *Journal of Human Capital 4 (4)*, 331–364.

Haider, S. and G. Solon (2006). Life-cycle variation in the association between current and lifetime earnings. *The American Economic Review 96*(4), 1308–1320.

Hertz, T. (2007). Trends in the intergenerational elasticity of family income in the United States. *Industrial Relations 46 (1)*, 22–50.

Hertz, T. N. (2001). Education, inequality andeconomic mobility in South Africa. Ph.D. thesis, University of Massachusetts.

Holmlund, H. (2006). Intergenerational mobility and assortative mating: Effects of an educational reform. working paper 4/2006, Swedish Institute for Social Research, Stockholm University.

Kalkbrenner, E. and E. Villanueva (2007). Intergenerational mobility in income and education in Spain. Mimeo.

Kremer, M. and E. Maskin (1995). Wage inequality and segregation by skill. NBER Working Paper num. 5718.

Lam, D. and R. F. Schoeni (1993). Effects of family background on earnings and return to schooling: evidence from brazil. *Journal of Political Economy 101*(4), 710–40.

Lee, C.-I. and G. Solon (2009). Trends in intergenerational income mobility. *Review of Economics and Statistics 91*(November), 766–772.

Leigh, A. (2007). Intergenerational mobility in Australia. *The B.E. Journal of Economic Analysis & Policy 7,*(Iss. 2 (Contributions)), Article 6.

Levine, D. I. and B. Mazumder (2007). The growing importance of family: Evidence from brothers' earnings. *Industrial Relations 46 (1)*, 7–21.

Levitt, S. and S. J. Dubner (2005). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. William Morrow/HarperCollins.

Lillard, L. A. and M. R. Kilburn (1995). Intergenerational earnings links: Sons and daughters. Papers 95-17, RAND - Labor and Population Program.

Long, J. and J. Ferrie (2011). Intergenerational occupational mobility in Britain and the U.S. since 1850. *American Economic Review, forthcoming*.

Marimon, R. and F. Zilibotti (1999). Unemployment vs. mismatch of talents: Reconsidering unemployment benefits. *Economic Journal 109*(455), 266–291.

Mayer, S. E. and L. M. Lopoo (2005). Has the intergenerational transmission of economic status changed? *Journal of Human Resources 40*(1), 169–85.

Miley, T. J. (2004). The politics of language and nation: The case of the Catalans in contemporary Spain. Ph.D. thesis, Department of Political Science at Yale University.

Ng, I. (2007). Intergenerational income mobility of young singaporeans. *The B.E. Journal of Economic Analysis & Policy 7,*(Iss. 2 (Topics)), Article 3.

Olivetti, C. and D. Paserman (2011). In the name of the father: Marriage and intergenerational mobility in the United States, 1850-1930. mimeo.

Osterbacka, E. (2001). Family background and economic status in Finland. *Scandinavian Journal of Economics 103*(3), 467– 84.

Osterberg, T. (2000). Intergenerational income mobility in Sweden: What do tax data show? *Review of Income and Wealth. 46*(4), 421–36.

Page, M. E. and G. Solon (2003). Correlations between brothers and neighboring boys in their adult earnings: The importance of being urban. *Journal of Labor Economics 21*, 831–55.

Parman, J. (2011). American mobility and the expansion of public education. *The Journal of Economic History 71*(1), 105–132.

Pascual, M. (2009). Intergenerational income mobility: The transmission of socio-economic status in spain. *Journal of Policy Modeling 31*, 835–846.

Rubinstein, Y. and D. Brenner (2011). Pride and prejudice: Using ethnic-sounding names and inter-ethnic marriages to identify labor market discrimination. Mimeo, LSE.

Shimer, R. and L. Smith (2000). Assortative matching and search. *Econometrica 2*(68), 343–369.

Solon, G. (1992). Intergenerational income mobility in the United States. *American Economic Review 82*(3), 393–408.

Solon, G. (1999). *Intergenerational Mobility in the Labor Market*. in Orley C. Ashenfelter and David Card (eds.), Handbook of Labor Economics, Volume 4B, Amsterdam: North-Holland, pp. 1761-1800.

Solon, G. (2002). Cross-country differences in intergenerational earnings mobility. *Journal of Economic Perspectives 16*(3), 59– 66.

Solon, G., M. Corcoran, Roger, and L. Deborah (1991). A longitudinal analysis of siblings correlations in economic status. *Journal of Human Resources 26*, 509–34.

Wiegand, J. (1997). Intergenerational earnings mobility in Germany. mimeo.

A Appendix: Proofs of Model's Properties

Proof of Property 1

Consider some date $t-1$ and a surname $s \in \Omega$ such that $F_{t-1}(s) > 0$. Define $Q_{t-1}(s)$ as the number of individuals with surname s so that $Q_{t-1}(s) = N_{t-1}F_{t-1}(s)$. Note that, conditional on $Q_{t-1}(s)$, $Q_t(s)$ is a binomial random variable with support $[0, m\,Q_{t-1}(s)]$ and distribution (suppressing the '(s)' notation)

$$\text{Prob}(Q_t = km) = \left(\begin{array}{c} Q_{t-1} \\ k \end{array} \right) q^k (1-q)^{Q_{t-1}-k} \quad . \tag{11}$$

The conditional mean and variance of Q_t are $Q_{t-1}mq$ and $Q_{t-1}m^2q(1-q)$, respectively. Therefore,

$$Q_t \;\; = \;\; Q_{t-1}mq + w_t \quad , \tag{12}$$

where w_t is defined as the innovation, $w_t \equiv Q_t - E_{t-1}Q_t$. If $mq = 1$ then Q_t follows a driftless random walk.[23]

Proof of Property 3

Fix some date t. Partition the population into *families*: groups of individuals who share the same lineage (which is possible because of asexual reproduction). Suppose, to begin with, that every family's lineage dates back k periods and that no two families share the same surname. Then the cross-sectional mean and variance of income for each family are, respectively,

$$E(y_{ist} \,|\, s) \;\; = \;\; \rho^k y_{s,t-k} \tag{13}$$

$$Var(y_{ist} \,|\, s) \;\; = \;\; V_\varepsilon \sum_{l=0}^{k-1} \rho^{2l} \tag{14}$$

where $y_{s,t-k}$ is the income of the patriarch of the family with surname s. Now recognize that the society-wide cross-sectional variance can be decomposed into the average within-family variance and across-family variance in the conditional mean:

$$Var(y_{ist}) \;\; = \;\; E\big(\, Var(y_{ist} \,|\, s)\big) + \, Var\big(E(y_{ist} \,|\, s)\big) \quad . \tag{15}$$

[23] This is related to the "branching process" literature. It was started by Francis Galton in 1873. He posed the following problem (our model with zero mutation).

> *Problem 4001:* A large nation, of whom we will only concern ourselves with the adult males, N in number, and who each bear separate surnames, colonise a district. Their law of population is such that, in each generation, a_0 percent of the adult makes have no make children who reach adult life; a_1 have one such male child; a_2 have two; and so on up to a_5 who have five.
>
> Find (1) what proportion of the surnames will have become extinct after r generations; and (2) how many instances there will be of the same surname being held by m persons.

The answer was finally figured out using martingale methods, but not until in 1950! It's kind of complicated, but the upshot is that, with strictly positive population growth a fraction q of all surnames with vanish with probability 1 and a fraction $(1-q)$ will persist forever (U.S. data on q suggests about 0.8). The distribution for the surviving names is exponential. This is from "Branching processes since 1873," by David Kendall. Google this title and you'll find it right away.

The ICS from equation (5) is proportional to the second term on the right which, according to expression (13) is monotonically increasing in ρ. This proves Property 3 for the case identical lineage horizons and unique within-family surnames.

Consider next the general case of lineage horizons that vary across families. Suppose that family j all derive from a patriarch who lived k_j periods before date t. Then equations (13) and (14) remain valid for each k_j and equation 15 takes the form of family-size weighted means and variances. Nevertheless, holding fixed the structure of the population, the second term on the right of equation (15) remains a monotonically increasing function of ρ.

Finally, relax the assumption that surnames and families are uniquely associated. If family j_1 and family j_2 share the same surname, s, then $E(y_{ist}\,;\,s)$ is a family-size-weighted average of the incomes of all of the members of the two families. Such averaging will, of course, decrease the cross-sectional variance in the conditional means, $Var(E(y_{ist}\,|\,s))$, thereby decreasing the ICS. However, holding fixed the population structure, it remains the case that this conditional variance, and the ICS, are increasing in ρ.

Proof of Property 4

Here, we demonstrate that our assortative mating model from Section 10.1 has a unique stationary distribution and derive expressions for the models variances and correlations in terms of its structural parameters.

Recall that male and female children in the i^{th} household with surname s at date t have income described by

$$y^m_{ist} = rz_{ip,t-1} + e^m_{ist} \quad ; \quad y^f_{ist} = rz_{ip,t-1} + e^f_{ist} \quad , \tag{16}$$

where $z_{ip,t-1}$ is the average income of these children's parents, who formed this household at date $t-1$, r is the *household* inheritance parameter and the innovations e are *i.i.d.* $\mathrm{N}(0, V_e)$. We now suppress the i and s notation (they are not needed here). Mating is described by

$$y^f_{pt} = \lambda y^m_{pt} + u_{pt} \quad ; \quad u_{pt} \sim N\left(0, V_u\right) \quad . \tag{17}$$

First, we guess that there exists a stationary distribution for z that has the form $\mathrm{N}(0, V_z)$. If so, then parental income at date $t+1$ — formed from the date t mating rule (17) — satisfies

$$z_{pt} \;=\; \left(y^m_{pt} + y^f_{pt}\right)/2 \;=\; \left((1+\lambda)y^m_{pt} + u_{pt}\right)/2 \tag{18}$$

where the first equation is just the definition of average parental income and the second applies the mating rule (17). Applying the inheritance process (16), we get

$$2z_{pt} \;\;=\;\; \left(rz_{p,t-1} + e^m_t\right)\left(1+\lambda\right) + u_{pt} \quad .$$

The variance of the distribution of z, then (if it exists), results from taking the unconditional variance of both sides and imposing stationarity:

$$V_z \;\;=\;\; \frac{(1+\lambda)^2 V_e + V_u}{4 - (1+\lambda)^2 r^2} \quad . \tag{19}$$

This gives V_z as a function of the structural parameters λ, V_e and r, and the variance of mating noise, V_u,

which is uniquely determined below.

Next, note that a stationary distribution for z implies that the income of male and female children have the same distribution (*i.e.*, by inspection of Equation (16)). Thus, we can write $y^m_{ist} \sim N(0, V_y)$ and $y^f_{ist} \sim N(0, V_y)$, for some variance, V_y, to be uniquely determined below. Given this, the mating rule, Equation (17), imposes that

$$V_u = (1-\lambda^2)V_y \ . \tag{20}$$

This guarantees that the distribution of female income implied by the mating rule coincides with that implied by the inheritance process.

Next, consider the income of males at date $t+1$. Using (16) and (18):

$$y^m_{t+1} = \frac{r(1+\lambda)}{2} y^m_{pt} + ru_{pt}/2 + e^m_{t+1}. \tag{21}$$

Since $r < 1$ and $\lambda < 1$, then $r(1+\lambda)/2 < 1$. Given the independence assumptions on u and e, and given that fertility is deterministic (with each male bearing one male offspring), then Equation (21) gives the income of a male as stationary Gaussian first-order autoregressive function of the income of his father. Its unconditional distribution is

$$y^m_t \sim \mathrm{N}\big(0\,,\, \frac{r^2 V_u/4 + V_e}{1 - r^2(1+\lambda)^2/4}\big) \ .$$

By a cross-sectional law-of-large numbers, this also gives the stationary cross-sectional distribution of male income.

All that remains is to solve for V_y as a function of the model's structural parameters. Using this last expression and rearranging:

$$V_y = \frac{V_e}{\lambda(1+\lambda)} \ ,$$

where the second equation follows from Equation (20) and the third follows from solving for V_y and rearranging. Substituting the result into Equation (19) yields:

$$V_z = \frac{\lambda(1+\lambda)^2 V_e + 2(1-\lambda)V_e}{\lambda(4-(1+\lambda)^2 r^2)}$$

This implies that there does indeed exist a stationary distribution for average parental income, z, that is consistent with the inheritance and mating rules, (16) and (17). The variance of the *male* inheritance shock, w^m_{ist} from Equation (10), is

$$V_w = V_e\big(1 + \frac{r^2(1-\lambda)}{4\lambda}\big).$$

B Appendix: Surnames as proxy of Ethnicity

How good of a proxy of ethnicity is our *CatalanDegree* variable? We know several things that can help us understand. First, a large percentage of intra-Spanish immigration into Catalonia occurred after 1955. Second, this immigration flow was large; without it Cabré (2004) estimates that the year 2000 population would have been 2.7 million instead of the actual value of roughly 6 million.

These facts tell us that, in the 2001 census, older people, and especially those born in Catalonia, are more likely to be of Catalan origin than younger people. If our *CatalanDegree* variable is a good proxy for regional origin, it should therefore reflect this. Table 7 shows that it does. The overall average of *CatalanDegree* in 0.34, whereas among people born prior to 1950, in Catalonia, the average is 0.57. Figure 13 elaborates. It plots the mean and standard deviation of *CatalanDegree* for the same rolling window of cohorts used in Figure 8. The surname distribution in Catalonia has clearly become 'less Catalan' over time, as the immigration flows tell us it should if our proxy is a good one.

As further support for the quality of our *CatalanDegree* proxy we run two probit regressions. In the first, the left-hand-side (LHS) variable takes value 1 if an individual has full knowledge of the Catalan language.[24] The right-hand-side variables (RHS) are, in column (1), individual-specific controls (place of birth and age dummies). In column (2) our *CatalanDegree* variable is added. Results are reported in Table 8(a). We estimate a large, significant, positive probability. Figure 14(a) shows the estimated probability for the relevant range of the *CatalanDegree* variable.

The second regression asks how well *CatalanDegree* predicts immigration history. The LHS variable takes value 1 if an individual 50 years of age or older immigrated into Catalonia from elsewhere in Spain. Results are reported in Table 8(b) and Figure 14(b). The estimates are negative, large and significant and the pseudo-R^2 increases dramatically with the inclusion of *CatalanDegree*. People with lower *CatalanDegree* surnames are much more likely to be immigrants than those with higher *CatalanDegree* surnames.

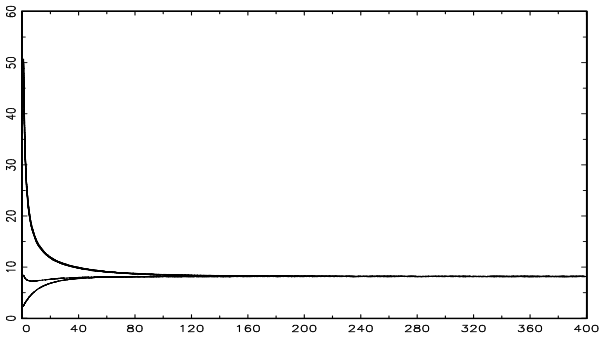
To sum up, the *CatalanDegree* variable seems to approximate ethnicity quite well.

C Invented Catalonias

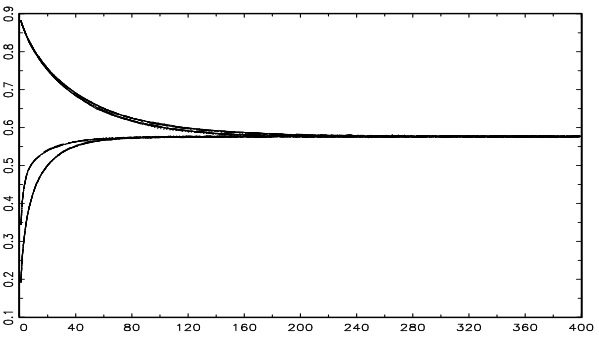
Our results should not be sensitive to any random (but sufficiently large) partitioning of the surnames set. As as example, we divide the letters of the alphabet to two groups, "first half" and "second half". Table 9 does exactly this. The first column reports, for comparison purposes, the overall ICS from Table 2. The second and third report the same statistics but for two "invented Catalonias:" those from the first half of the alphabet and those from the second half.[25] As our model predicts, neither the R^2 of the regressions nor the ICS change across the populations.

[24] The census question asks a resident if she speaks, reads and writes Catalan. Roughly 45% of the over-25 population responded in the affirmative.

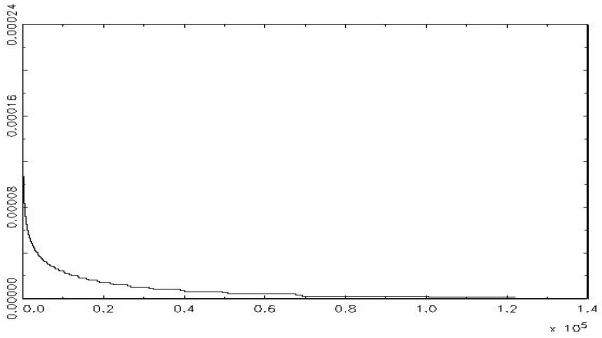
[25] We have done the experiment with other random groupings, and obtained the same result.



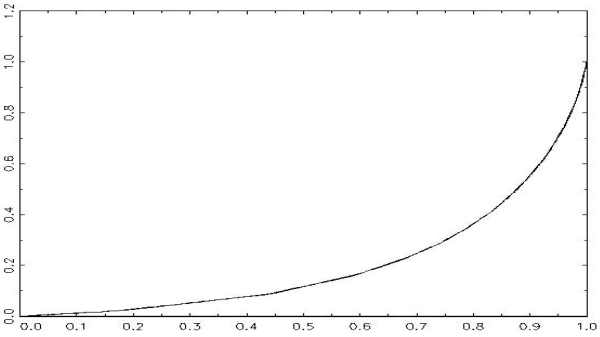
(a) Time series of average number of agents per surname



(b) Times series of Gini of surname distribution



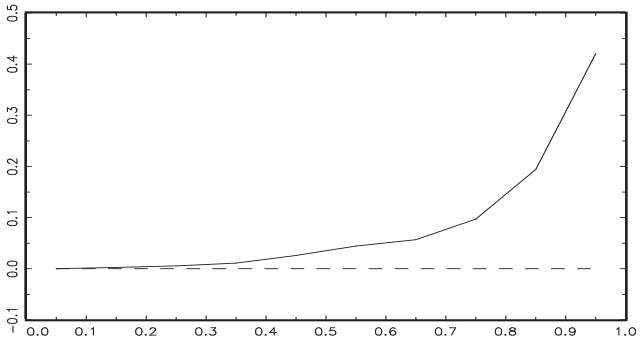
(c) Ordered frequency distribution, $G(k)$



(d) Lorenz curve of surname distribution

Figure 1: Time series of number of agents per surname and Gini coefficient, and, for different values of ρ, the surname distribution $G(k)$ and associated Lorenz curve.

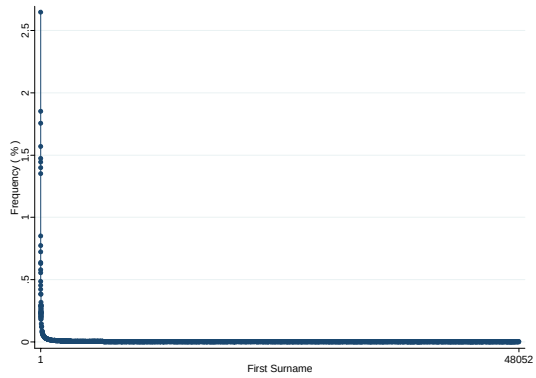
Notes: Model Simulations with baseline parameter values: N_0=1000000; V_ε=1.000; μ=0.0200; q=0.50; m=2; $\rho \in [0.05, 0.95]$. Different initial conditions: number of surnames= 10, 1000, 100000 and 1000000.



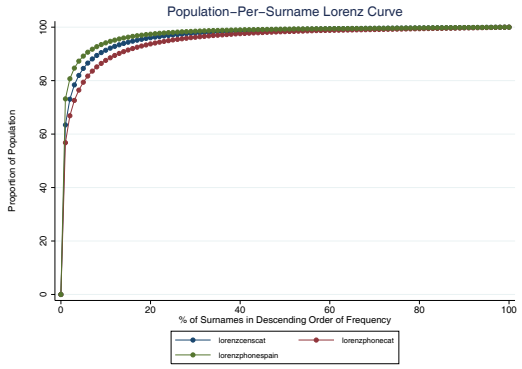
(a) Average R^2_L (solid line) and R^2_F (dotted line) against ρ

Figure 2: Surnames are informative, and their informational content increases with the degree of inheritance that there is in society.

Notes: Model Simulations with baseline parameter values: as in Figure 1.



(a) Distribution of First Surname



(b) Lorenz Curve of the Surname Distribution in Catalonia and Spain

Figure 3: Distribution of the first surname in Catalonia and Lorenz curves for Spain and Catalonia

For 3(a): Source: 2001 Catalan census. Population: Spanish citizens living in Catalonia aged 25 and above, all surnames.
For 3(b): Source: 2004 Spanish telephone directory and 2001 Catalan census. Population: All phones with first and second surnames not missing. Population percentage per surname (1% Steps).

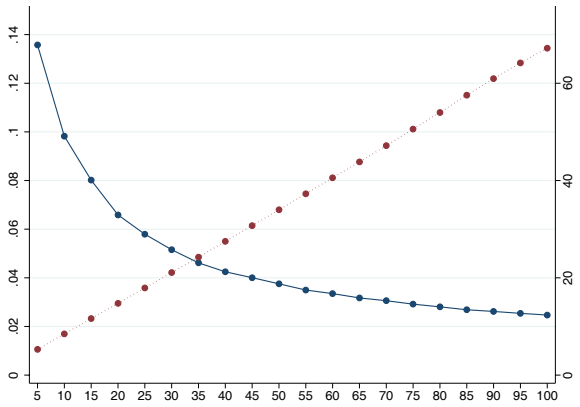
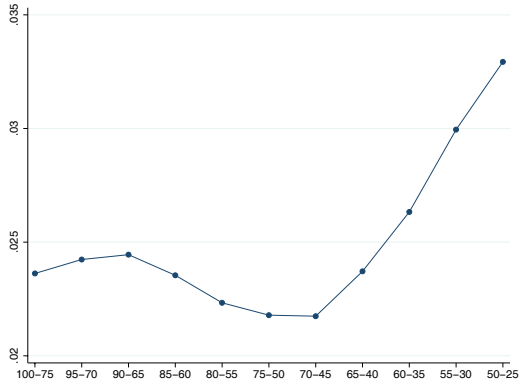
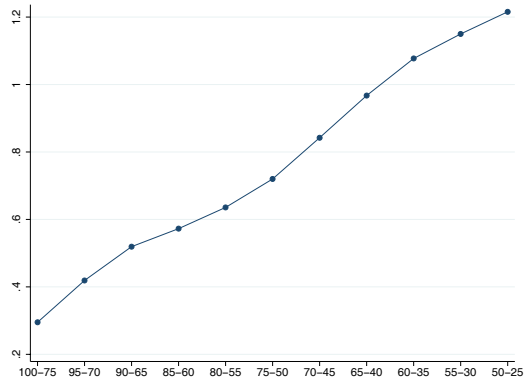


Figure 4: ICS is larger for less frequent surnames

Notes: ICS (solid line) and individuals per surname (dotted line). Regressions as in Table 2 (Columns 3 and 4) by percentiles, where percentile x corresponds to the x% least-frequent surnames. Source: 2001 Catalan Census.



(a) Evolution of ICS



(b) Evolution of parameter of *CatalanDegree*

Figure 5: Evolution of ICS and parameter of *CatalanDegree* over moving windows of cohorts

Notes: Regressions as in Table 2 (Columns 3 and 4). The overlapping sequence of cohorts starts with those aged 75-100 years old in 2001, then continues with those aged 70-95 years old, and so on, ending with the 25-50 year-old cohort. Source: 2001 Catalan Census.

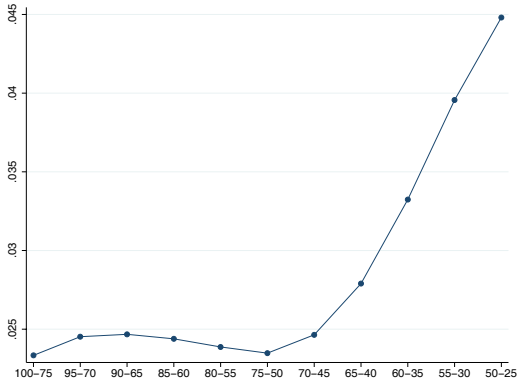
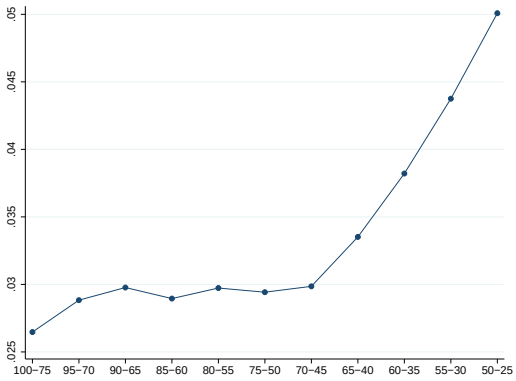
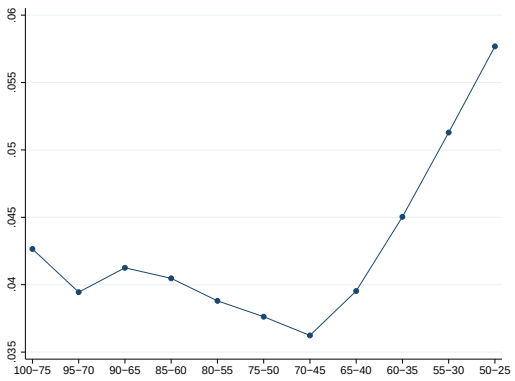


Figure 6: Evolution of ICS over moving windows of cohorts. No ethnic controls.

Notes: Regressions as in Table 2 (Columns 5 and 6). Overlapping age-cohorts are described in caption to Figure 5. Source: 2001 Catalan Census.



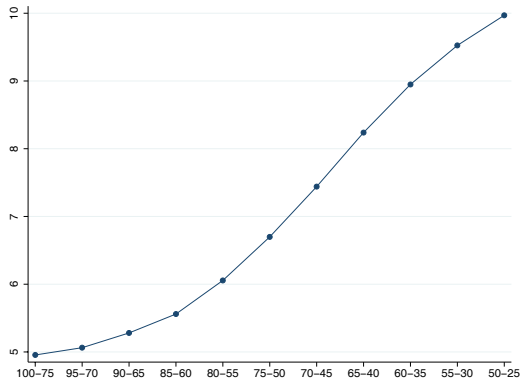
(a) 50% Most Catalan Surnames



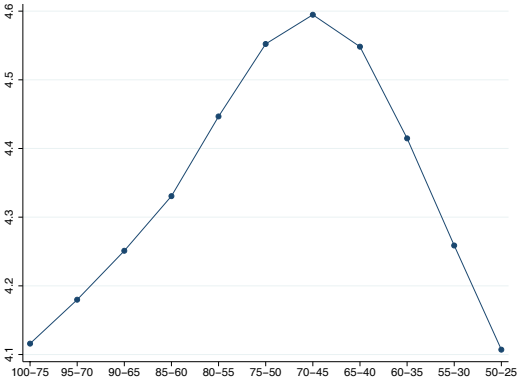
(b) 50% Least Frequent Surnames

Figure 7: Evolution of ICS over moving windows of cohorts, subpopulations.

Notes: Regressions as in Table 2 (Columns 5 and 6). Overlapping age-cohorts are described in caption to Figure 5. Source: 2001 Catalan Census.



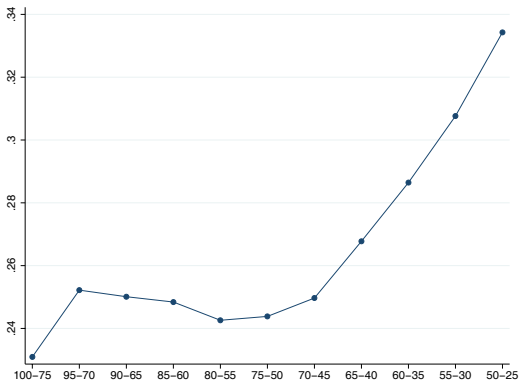
(a) Average



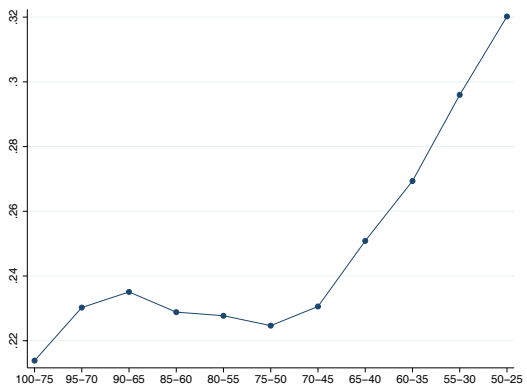
(b) Standard Deviation

Figure 8: Evolution of years of education over moving windows of cohorts

Notes: Overlapping age-cohorts are described in caption to Figure 5. Source: 2001 Catalan Census.



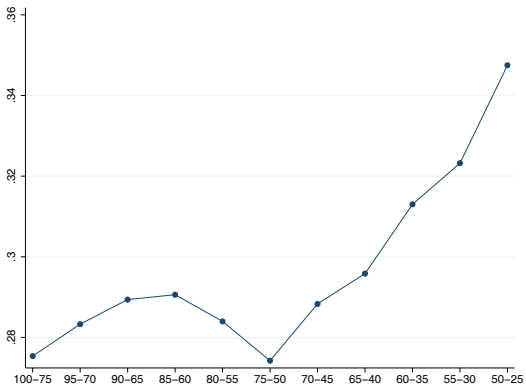
(a) Complete surnames shared by 2 people



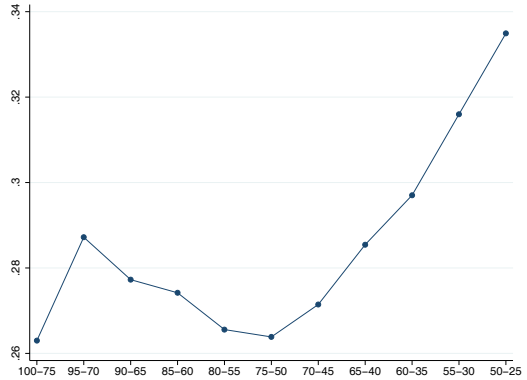
(b) Complete surnames shared by 2 or 3 people

Figure 9: Evolution of Sibling Correlations, ICSIB over moving windows of cohorts.

Notes: Regressions as in Table 6(a). Overlapping age-cohorts are described in caption to Figure 5. Source: 2001 Catalan Census.



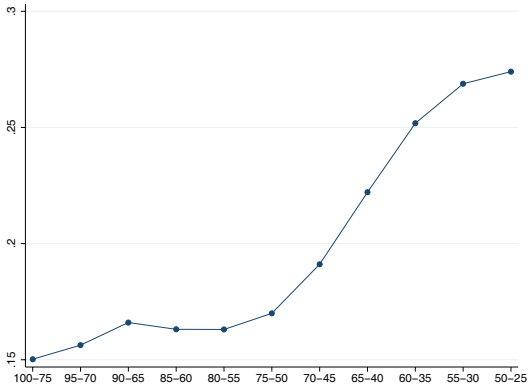
(a) Complete surnames shared by 2 people



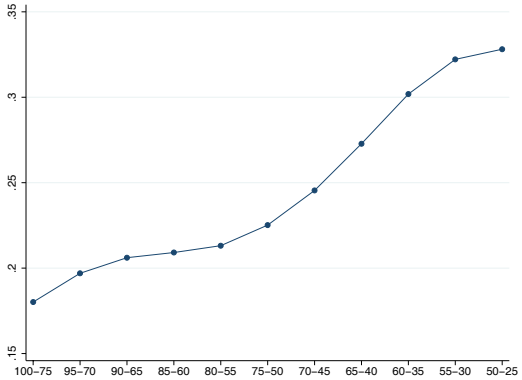
(b) Complete surnames shared by 2 or 3 people

Figure 10: Evolution of Sibling Correlations, ICSIB moving windows of cohorts. 50% Most Catalan Surnames

Notes: Regressions as in Table 6(b). Overlapping age-cohorts are described in caption to Figure 5. Source: 2001 Catalan Census.



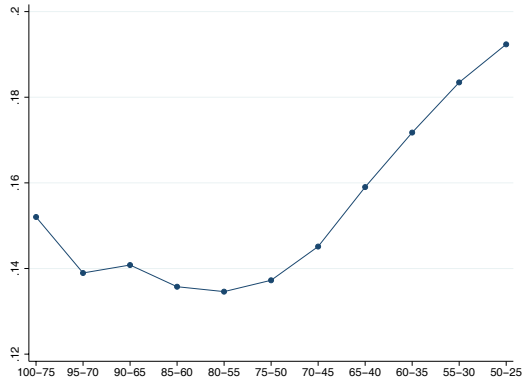
(a) AM in Education



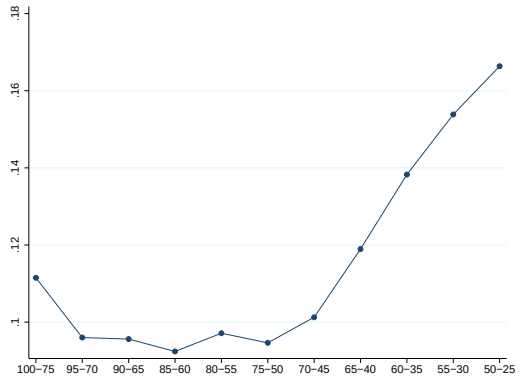
(b) AM in *CatalanDegree*

Figure 11: Evolution of Assortative Mating in Education & *CatalanDegree* over moving windows of cohorts.

Notes: Regressions include age and place of birth dummies. Overlapping age-cohorts are described in caption to Figure 5. Populations: Spanish citizens living in Catalonia with frequency of first and second surname larger than one. Source: 2001 Catalan Census.



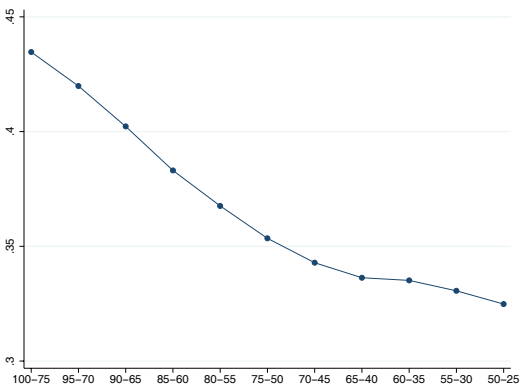
(a) 50% Most Catalan Surnames



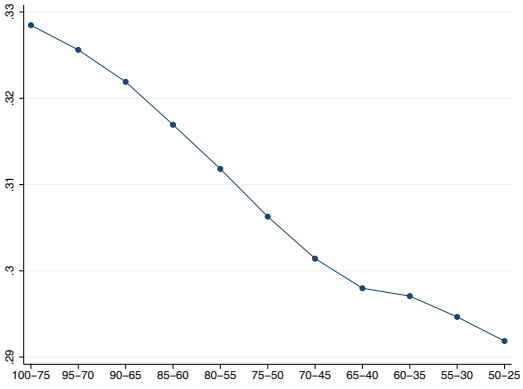
(b) 50% Least Frequent Surnames

Figure 12: Evolution of Assortative Mating in Education over moving windows of cohorts, sub-populations.

Notes: Regressions as in Figure 11. Overlapping age-cohorts are described in caption to Figure 5. Source: 2001 Catalan Census.



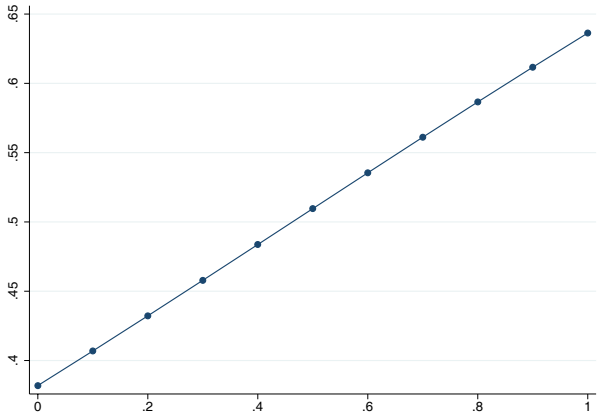
(a) Average of *CatalanDegree*



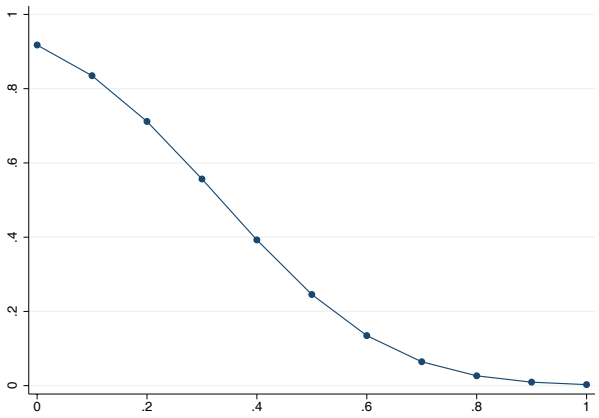
(b) Standard Deviation of *CatalanDegree*

Figure 13: Evolution of *CatalanDegree* over moving windows of cohorts

Notes: Overlapping age-cohorts are described in caption to Figure 5. Source: 2001 Catalan Census.



(a) Probability of Catalan language knowledge



(b) Probability of being an immigrant

Figure 14: Probabilities of Catalan language knowledge and of being an immigrant, as a function of *CatalanDegree*.

Notes: Regressions as in Table 8. For Figure 14(a), reference individual is a male, aged 50-55, born in the county of Barcelona. For Figure 14(b), reference individual is a male, aged 60-65. Source: 2001 Catalan Census.

Table 1: Surnames Distribution: Gini Index and People per Surname in Catalonia and Spain

	Spain		Catalonia	
	(PhoneBook)	(Census)	(PhoneBook)	(Census)
	(1)	(2)	(3)	(4)
	All	All	All	Only males
Number of People	11,397,116	6,123,909	2,073,219	2,983,384
Number of Surnames	155,782	91,568	61,396	63,141
People per Surname	73.161	66.878	33.768	47.249
Gini Index	0.9485	0.9304	0.9028	0.908

Source: 2004 Spanish Phone Book & 2001 Catalan Census. Populations: Columns (1-3): All individuals/phones with first & second surnames not missing. Column 4: Men with first & second surnames not missing.

Table 2: ICS. Baseline population.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.706(0.011)	1.015(0.012)	1.707 (0.011)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.2652	0.2735	0.2980	0.2735	0.2955	0.2653
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.534	0.000	0.601

Notes: All regressions include age and place of birth dummies. Fake-surnames have the same distribution as Surnames and are allocated randomly. (*) F-test if Surname dummies are jointly significant. Standard errors in parenthesis. The very small standard errors of our estimates are obviously due to a very large sample size. As a result, most of the discussion in the text upon the economic magnitude of the coefficients, not their statistical significance. One important exception is the joint significance of the actual versus fake surname dummies. Population: Spanish citizens living in Catalonia aged 25 and above, with frequency of first surname larger than one. Number of observations: 2,057,134. Number of surnames: 30,610. Source: 2001 Catalan Census.

Table 3: ICS. Subpopulations.

(a) Born in Catalonia.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.683 (0.012)	0.98(0.013)	1.682(0.012)		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.1543	0.1668	0.2016	0.1666	0.1979	0.1541
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.934	0.000	0.917

(b) Born in Catalonia before 1950. ("Old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.019(0.021)	0.609(0.022)	1.017(0.021)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.1331	0.1375	0.1752	0.1373	0.1737	0.1329
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.698	0.000	0.687

(c) 50% Most Catalan Surnames.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		0.971(0.014)	0.783(0.015)	0.972(0.014)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.2466	0.2501	0.2777	0.2501	0.2757	0.2467
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.299	0.000	0.318

Notes: Regressions as in table 2. For 3(a): Number of observations: 1,328,003. Number of surnames: 28,523. For 3(b): Number of observations: 465,896. Number of surnames: 20,793. For 3(c): Number of observations: 1,028,567. Number of surnames: 23,892. Source: 2001 Catalan Census.

Table 4: ICS. 50% Least Frequent Surnames.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.467(0.015)	0.801(0.016)	1.464(0.015)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.2597	0.2664	0.3038	0.2666	0.3020	0.2600
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.12	0.000	0.095

Notes: Regressions as in table 2. Number of observations: 1,028,727. Number of surnames: 30,275. Source: 2001 Catalan Census.

Table 5: Calibration of Baseline Model

Parameter	Value
ρ	0.567
μ	0.00856
m	6

Targets	Model	Data
ICS	0.03	0.0302
Gini Surnames	0.754	0.908
People per Surname	44.855	47.250

Notes: Parameters set to minimize a weighted distance to the targets. Found by searching in a grid where each of the parameter configurations was run for a population of 1.5 million individuals. Each configuration was run for 125 periods to get into steady state, and then the moments of the resulting economy were calculated and averaged for 25 further periods. In all configurations V_ϵ is set so that the standard deviation of the unconditional education distribution equals its data value, 4.655. The Gauss programs used for simulating the economies and for finding the parameter configuration that approximates the most to the data available from the authors upon request.

Table 6: Sibling Correlations, ICSIB. Rare complete–surnames.

(a) Spanish citizens living in Catalonia

LHS: years of education	(1)	(2)
Adjusted R^2, Complete–Surname Dummies	0.5025	0.4884
Adjusted R^2, Complete–Fake–Surnames Dummies	0.2517	0.2557
Informational Content of Siblings (ICSIB)	0.2508	0.2327
Observations	428,134	655,303
Number of Complete–Surnames	214,067	289,790
Max number of People per Complete–Surname	2	3

(b) 50% Most Catalan Surnames

LHS: years of education	(1)	(2)
Adjusted R^2, Complete–Surname Dummies	0.5029	0.4904
Adjusted R^2, Complete–Fake–Surnames Dummies	0.2446	0.2434
Informational Content of Siblings (ICSIB)	0.2583	0.2470
Observations	302,486	453,219
Number of Complete–Surnames	151,243	201,489
Max number of People per Complete–Surname	2	3

Notes: All regressions include age and place of birth dummies. Fake–complete–surnames have the same distribution as complete–surnames and are allocated randomly. For 6(a): Population: Spanish male citizens living in Catalonia aged 25 and above, with frequency of complete–surname larger than one. For 6(b): Population: Individuals with 50% most Catalan complete–surname of sample in table 6(a). Source: 2001 Catalan Census.

Table 7: *CatalanDegree* Summary Statistics

	Complete population of males	Born in Catalonia before 1950	Born anywhere in Spain before 1950	Born anywhere in Spain after 1950
	(1)	(2)	(3)	(4)
Mean CatalanDegreeSurname2	0.344	0.566	0.367	0.324
Standard deviation	(0.302)	(0.324)	(0.312)	(0.292)
Share with CatalanDegreeSurname2>0.16	0.568	0.836	0.596	0.545

Notes: Column (1): Male Spanish citizens living in Catalonia aged 25 and above, with frequency of first surname larger than one, number of observations: 2,057,134. Column (2): Individuals born in Catalonia before 1950 of sample in column (1), number of observations: 465,896. Column (3): Individuals born before 1950 of sample in column (1), number of observations: 937,441. Column (4): Individuals born after 1950 of sample in column (1), number of observations: 1,119,693. Source: 2001 Catalan Census.

Table 8: *CatalanDegree* & Probabilities of Catalan language knowledge and of being an immigrant

(a) Probability of Catalan language knowledge

LHS: Knowledge of Catalan	(1)	(2)
CatalanDegreeSurname2		0.649 $_{(.004)}$
Log likelihood	-1106700.3	-1092203.9
Pseudo R^2	0.2196	0.2298

(b) Probability of being immigrant

LHS: Immigrant	(1)	(2)
CatalanDegreeSurname2		-4.156$_{(.009)}$
Log likelihood	-645813.49	-410908.05
Pseudo R^2	0.0061	0.3676

Notes: Probit Estimates. All regressions include age dummies. Regressions in table 8(a) also include place of birth dummies. Standard errors in parenthesis. For 8(a): Sample: baseline population. The LHS variable *Knowledge of Catalan* takes value 1 for individuals who understand, can speak, can read and can write the Catalan language and zero otherwise. Number of observations: 2,057,134. For 8(b): Sample: Individuals born before 1950 of baseline population. The LHS variable *Immigrant* takes value 1 for individuals who were not born in Catalonia and zero otherwise. Number of observations: 937,441. Source: 2001 Catalan Census.

Table 9: ICS. "Invented" Catalonias.

LHS: years of education	(1)	(2)	(3)
CatalanDegreeSurname2	Yes	Yes	Yes
Adjusted R^2, Surname Dummies	0.2980	0.2992	0.2969
Adjusted R^2, Fake-Surname Dummies	0.2735	0.2728	0.2743
ICS	0.0245	0.0264	0.0226
Sample	All surnames	"First letters"	"Last letters"
Observations	2,057,134	1,046,996	1,010,138

Notes: Regressions as in table 2. (columns 3 and 4). Populations: Column (1): baseline population; Columns (2) and (3) are the first and second half respectively of the baseline population alphabetically ordered. Source: 2001 Catalan Census.